UNIVERSIDADE FEDERAL DO PARANÁ

LUIS JOSÉ ROHLING

EVIDÊNCIAS DA FALHA DO MODELO DE GILBERT-ELLIOTT PARA PERDA DE PACOTES EM REDES WIFI E ANÁLISE DE ALTERNATIVAS

CURITIBA

2017

LUIS JOSÉ ROHLING

EVIDÊNCIAS DA FALHA DO MODELO DE GILBERT-ELLIOTT PARA PERDA DE PACOTES EM REDES WIFI E ANÁLISE DE ALTERNATIVAS

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica, Área de Concentração Telecomunicações, Departamento de Engenharia Elétrica, Setor de Tecnologia, Universidade Federal do Paraná, requisito parcial para a obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Carlos Marcelo Pedroso

CURITIBA
2017

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
Setor TECNOLOGIA
Programa de Pós Graduação em ENGENHARIA ELÉTRICA
Código CAPES: 40001016043P4

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ENGENHARIA ELÉTRICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **LUIS JOSE ROHLING**, intitulada: "**Evidências da Falha do Modelo de Gilbert-Elliott para Perda de Pacotes em Redes WIFI e Análise de Alternativas** ", após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.
A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 18 de Agosto de 2017.

CARLOS MARCELO PEDROSO
Presidente da Banca Examinadora (UFPR)

EDUARDO PARENTE RIBEIRO
Avaliador Interno (UFPR)

EVELIO MARTÍN GARCÍA FERNÁNDEZ
Avaliador Interno (UFPR)

MAURO SERGIO PEREIRA FONSECA
Avaliador Externo (UTFPR)

RESUMO

As redes WiFi representam uma das formas de acesso à Internet mais utilizadas, sendo o estudo do comportamento das perdas de pacotes nestas redes de fundamental importância para a evolução e aplicação desta tecnologia. Porém, o modelo clássico de Gilbert-Elliott, amplamente utilizado para descarte de pacotes em redes de dados, não é necessariamente o modelo mais adequado para os diversos cenários de utilização das redes WiFi, devido ao método de controle de acesso ao meio empregado nestas redes. Para caracterizar a perda de pacotes nestas redes, este trabalho apresenta o resultado de diversas medições de erros em redes WiFi. Estas medições foram feitas a partir do envio de um tráfego de controle, registrando-se também outros parâmetros importantes da rede sob teste. Os dados de perda de pacotes do fluxo de controle foram inicialmente tratados com um filtro passa-baixa e então aplicado um processo de clusterização para obtenção dos estados da rede. Na etapa seguinte, foi realizada a análise estatística dos estados BOM e RUIM da rede, identificando-se uma dependência temporal de longa duração entre o tempo de duração de estados subsequentes, o que indica que o modelo de Gilbert-Elliott não se aplica nestes cenários. Finalmente, é apresentada uma alternativa para a modelagem de perdas de pacotes em redes WiFi considerando os dados amostrais. Como alternativa, foi analisada a aderência ao modelo *Fractional Auto Regressive Integrated Moving Average* (FARIMA) para o tempo de duração dos estados BOM e RUIM. Os resultados indicam que o modelo FARIMA não possui boa aderência aos dados empíricos observados, mas a análise realizada mostrou quais os motivos para a não aderência e indica quais os próximos passos nesta pesquisa.

Palavras-chave: Redes sem fio. Perdas de pacotes. Modelo de Gilbert-Elliott.

ABSTRACT

WiFi networks are widely used for Internet access. Therefore the study of packet loss in WIFi networks is very important for the evolution and application of this technology. The classic Gilbert-Elliott model for packet loss, widely used in data network, does not capture the behavior of packet loss observed in WiFi networks, mainly because the medium access control used to share the channel among the wireless devices. This work presents the result of several samples of errors in WiFi networks. The samples was made by sending a control flow, and registering the performance metrics as the packet loss, network utilization, signal-to-noise ratio, among others. The packets of control flow were used to identify the GOOD and the BAD states using a clustering method. In the next step, the statistical analysis of the duration of GOOD and BAD states of the network was performed, identifying a temporal dependence, which indicates that the Gilbert-Elliott model does not apply in these scenarios. Finally, alternative models are presented. A long range dependence is observed. The Fractionally Auto Regressive Integrated Moving Average (ARIMA) was analyzed to modeling the duration of the GOOD and BAD States. The results indicate that the FARIMA does not fit perfectly the empirical data, however, it’s a good start for a new model.

Key words: Wireless networks. Packet loss. Gilbert-Elliott model.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

ACF	*Autocorrelation Function*
AP	*Access Point*
AR	*Autoregressive*
ARIMA	*Autoregressive Integrated Moving Average*
CSMA/CA	*Carrier-Sense Multiple Access with Collision Avoidance*
DELT	Departamento de Engenharia Elétrica
FARIMA	*Fractional Autoregressive Integrated Moving Average*
FIFO	*First In First Out*
GE	Gilbert-Elliott
GMLE	*Gaussian Maximum Likelihood Estimation*
HMM	*Hidden Markov Model*
IEEE	*Institute of Electrical and Electronics Engineers*
IoT	*Internet of Things*
ITU	*International Telecommunication Union*
LLCD	*Log-Log Complementary Distribution*
MA	*Moving Average*
PSR	Priestley-Subba Rao
UFPR	Universidade Federal do Paraná
WiFi	*Wireless Fidelity*
6LoWPAN	*IPv6 over Low power Wireless Personal Area Networks*

LISTA DE SÍMBOLOS

α	Parâmetro de forma
β	Parâmetro de localização
$\rho(k)$	Função de auto correlação
μ	Média
σ	Desvio padrão
$\emptyset_n$	Coeficiente do modelo auto regressivo (AR)
θ_n	Coeficiente do modelo de média móvel (MA)
δ	Fator de amortização do filtro digital passa-baixa
k	Deslocamento ou lag

SUMÁRIO

1 INTRODUÇÃO

O padrão IEEE 802.11, também chamado de WiFi, tem sido a principal opção utilizada para implementação de redes de acesso nos ambientes onde existe mobilidade de usuários. Estes ambientes compreendem principalmente as áreas de grande circulação de pessoas, onde empresas e governo buscam uma solução para disponibilizar o acesso à internet. Neste cenário, deve-se considerar também o crescente número de dispositivos conectados através dessa tecnologia de rede de acesso, e que deverá sofrer um acréscimo considerável com a nova onda chamada de Internet das Coisas (IoT – *Internet of Things*), que envolve diversas tecnologias, protocolos e aplicações (AL-FUQAHA et al., 2015). Assim, é de fundamental importância entender o comportamento destas redes, de forma a prever o seu desempenho e permitir o dimensionando dos recursos necessários. Um dos parâmetros críticos a ser considerado é a ocorrência de erros na transmissão, que se reflete na perda de pacotes, impactando significativamente no desempenho e na percepção do usuário quanto à qualidade da rede.

Ao longo do tempo, diversos modelos de perda de pacotes foram desenvolvidos, associadas à amostragem do tráfego real em diversas tecnologias de redes. Porém, com a mudança do perfil de utilização destas redes, aliado ao surgimento de novas tecnologias e aumento do tráfego, existe a necessidade de uma constante análise destes modelos.

As principais causas de perdas de pacotes são os fatores físicos, como ruído e desvanecimento do canal, e por fatores intrínsecos da tecnologia empregada, tal como o controle de acesso ao meio, quando dois ou mais equipamentos concorrem pelo uso da rede de rádio.

Entre os modelos de perdas de pacotes disponíveis destaca-se o proposto por Gilbert-Elliott que utiliza uma cadeia de Markov de dois estados para tentar capturar as características de rajada observadas na perda de pacotes (HASSLINGER; HOHLFELD, 2008). Porém, devido à concorrência no acesso ao meio utilizada atualmente pelo protocolo IEEE 802.11, o modelo de Gilbert-Elliott pode não ser o mais adequado.

1.1 Objetivos

O objetivo geral desta dissertação é evidenciar que o modelo Gilbert-Elliott não é o mais adequado para perdas de pacotes em redes sem fio padrão 802.11 e apresentar as possíveis alternativas.

Objetivos específicos:

- Realizar a amostragem de dados reais;
- Identificar os estados da rede utilizando métodos de clusterização;
- Verificar a aderência do modelo Gilbert-Elliott aos dados amostrados;
- Analisar os dados amostrados para identificar características associadas às distribuições estatísticas;
- Apresentar alternativas para modelagem de perdas de pacotes considerando os dados amostrados.

Os objetivos desta dissertação são parte do projeto de doutorado de Carlos Alexandre Gouvêa da Silva, que contribuiu na condução dos experimentos e na análise dos resultados.

1.2 Estrutura da dissertação

O Capítulo 2 apresenta os fundamentos teóricos envolvidos no estudo realizado. São descritos o modelo de Gilbert-Elliott, que é o modelo mais utilizado atualmente, bem como os principais modelos de perda apresentados na literatura. Também é apresentado o modelo de clusterização de dados utilizado para identificação dos surtos de erro, bem como os conceitos matemáticos necessários para a compreensão do modelo proposto e da análise de resultados. No Capítulo 3 são descritos os recursos utilizados e os métodos empregados na coleta, tratamento e análise dos dados. O modelo proposto é apresentado no Capítulo 4, bem como sua validação e comparações. As considerações finais, conclusões e identificação de possíveis trabalhos futuros são apresentados no Capítulo 5.

2 CONCEITOS FUNDAMENTAIS

Neste capítulo são apresentados os conceitos fundamentais necessários para compreensão do presente trabalho, incluindo o modelo Gilbert-Elliott e demais modelos de perdas de pacotes, os fundamentos matemáticos utilizados na modelagem de dados e métodos de clusterização de dados.

2.1 Perdas de pacotes em redes WiFi

O modelo apresentado por Gilbert (GILBERT, 1960) e Elliott (Elliott, 1963) é uma das principais abordagens utilizadas atualmente para a modelagem de perdas de pacotes em redes (HASSLINGER; HOHLFELD, 2008). Este modelo utiliza uma cadeia de Markov de 2 estados, denominados BOM (B) e RUIM (R), com probabilidade de incidência de erros pequena e grande, respectivamente. O modelo de cadeias de Markov é aplicado para fenômenos que não apresentam memória, ou seja, a transição para os estados BOM e RUIM depende apenas do estado atual e não dos anteriores.

A recomendação G.1050 da ITU-T (ITU-T, 2011) sugere modelos para avaliar o desempenho de transmissão multimídia sobre o protocolo IP. Esta recomendação utiliza o modelo de Gilbert-Elliott para a perda de pacotes em redes. Também é apresentada uma parametrização típica do modelo para o acesso, redes locais e núcleo da Internet com diversos níveis de qualidade.

No modelo de Gilbert-Elliott a probabilidade de erro depende do estado atual, com uma probabilidade de perda de $(1-k)$ para o estado B e $(1-h)$ para o estado R. A Figura 1 ilustra a transição de estados do modelo. A probabilidade para a troca do estado B para R é dado por *p* e a probabilidade para a troca do estado R para o B é dada por *r*.

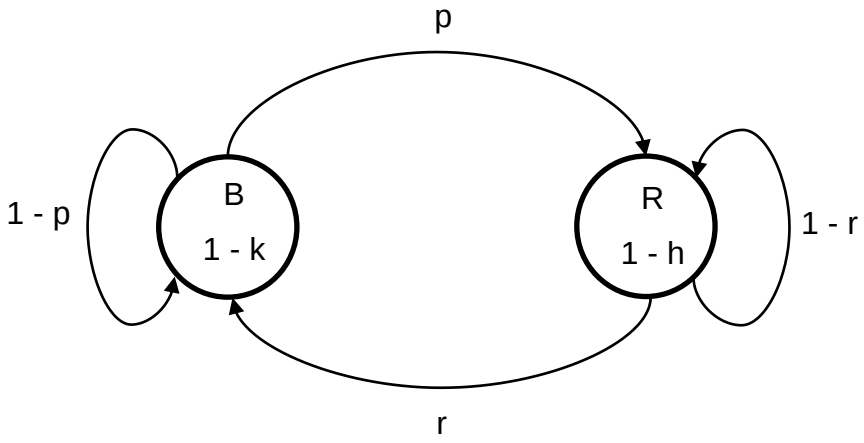


Figura 1 – Modelo de Gilbert-Elliott de 2 estados

A probabilidade de estado estacionário para este modelo é obtida aplicando-se o teorema do limite das cadeias de Markov (GRINSTEAD; SNELL, 2007), onde a probabilidade de perda será dada por $r(1-k)/(p+r)$ para o estado B e $p(1-h)/(p+r)$ para o estado R.

Gilbert propõe um modelo para caracterizar um canal de rajadas com ruído, adicionando memória ao canal simétrico binário codificado em dois estados da cadeia de Markov. Gilbert considera o caso especial de um estado BOM livre de erros (*k* = 1) e sugeriu, para estimar os parâmetros, um modelo de três instâncias mensuráveis de um processo de erro binário designado por $\{E_t\}$, onde $t \in \mathrm{N}$, e neste processo $\{E_t\} = 1$ indica um erro. As três instâncias definem a probabilidade de erro no primeiro passo, representada por a, a probabilidade de erro no primeiro e segundo passo, representada por b, e a probabilidade de existir um erro entre dois estados de erro, representada por c, e são expressas por:

$$a = \mathrm{P}(1), \;\; b = \mathrm{P}(1|1) \;\; \mathrm{e} \;\; c = \frac{\mathrm{P}(111)}{\mathrm{P}(101) + \mathrm{P}(111)} \qquad (1)$$

Conhecendo a, b e c, os três parâmetros do modelo podem ser calculados da seguinte forma:

$$r = 1 - \frac{a.c - b^2}{2.a.c - b.(\mathrm{a} + c)}; \quad h = 1 - \frac{b}{1-r}; \;\; p = \frac{a.r}{1-h-a} \qquad (2)$$

Gilbert argumenta que a medição de *c* pode ser evitada escolhendo $h = 0.5$ e usando $1 - r = 2b$. Além disso, ele mostrou que este método pode levar à parâmetros inconsistentes ($p, r, h < 0$ ou $p, r, h > 1$) se a observação for muito pequena. A Tabela 1 apresenta o resumo dos modelos, seus parâmetros e a complexidade de sua estimação.

Tabela 1 – Características dos Modelos de Gilbert e Elliott

Modelo	Parâmetro	Complexidade	Simplificação
Gilbert (simples)	*p, r*	Simples	$k = 1, h \in \{0, 0.5\}$
Gilbert	*p, r, h*	Média	$k = 1$
Gilbert - Elliott	*p, r, h, k*	Alta	

FONTE: (HASSLINGER; HOHLFELD, 2008)

A cadeia de Markov pode ser ampliada para mais estados, o que torna o modelo de Gilbert-Elliott mais complexo, exigindo maiores recursos computacionais para sua resolução. Estes modelos são tipicamente empregados em sistemas de multiplexação única, tais como várias fontes sendo atendidas por uma fila FIFO. Quanto maior a ordem da cadeia de Markov adotada, maior é a precisão do modelo quando comparado com o resultado do sistema simulado (YU et al., 2008).

O modelo de Gilbert-Elliott, porém, por ser baseado em cadeias de Markov, pode não apresentar um resultado adequado, dependendo das condições e natureza dos fenômenos subjacentes, como quando existe dependência temporal entre o tempo de duração de estados consecutivos. A distribuição de probabilidade do tempo de duração dos estados bom e ruim deveria apresentar uma distribuição do tipo exponencial para estar aderente ao modelo de Gilbert-Elliott. Porém, a característica de transmissão do meio físico das redes WiFi não necessariamente apresenta este comportamento, pois um congestionamento de rede, que tende a aumentar a taxa de perda de pacotes, pode causar um efeito de memória em virtude do algoritmo de controle de acesso ao meio.

O método de controle de acesso ao meio utilizado no padrão 802.11 é chamado de CSMA/CA (*Carrier Sense Multiple Access with Congestion Avoidance*) onde a estação que deseja transmitir primeiramente monitora o meio, verificando se

existe algum tráfego no canal utilizado (IEEE, 1997). Caso o canal esteja livre por um tempo maior que o definido para o chamado DIFS (*Distributed Interframe Space*) ela inicia a transmissão. Caso o canal esteja ocupado, a estação aguarda a liberação do canal pelo tempo mínimo do DIFS e inicia uma contagem de tempo aleatória, chamado de *backoff*, antes de iniciar a transmissão. Como o processo de transmissão não permite a detecção de colisão, a estação deverá receber um ACK enviado pelo receptor, confirmando a recepção do quadro sem erros. Caso o ACK não seja recebido dentro de um limite de tempo, a estação deverá retransmitir o quadro, seguindo o processo de monitoração e temporização. Caso ocorram sucessivas tentativas de transmissão sem sucesso, o intervalo de tempo considerado para a geração do tempo aleatório de espera vai aumentando, refletindo o congestionamento da rede. No outro método empregado no padrão 802.11, a estação não transmite imediatamente o quadro, após a detecção do canal livre, mas envia uma solicitação RTS (*Request to Send*) e aguarda a confirmação CTS (*Clear to Send*). Este modo permite que em uma rede operando no modo infraestrutura o AP gerencie a transmissão das diversas estações associadas a ele. Porém, poderão existir estações que conseguem identificar o tráfego gerado pelo AP mas não percebem o tráfego das estações mais distantes, gerando colisão, pois identificam o canal como disponível quando na realidade não estão. No modo de reserva de recursos, todas as estações receberão o CTS, contendo o tempo de transmissão permitido para uma determinada estação.

Porém, mesmo com o mecanismo CSMA/CA, colisões ainda podem ocorrer, e a probabilidade de colisão aumenta à medida que mais estações se associem ao AP, levando à um provável aumento na perda de pacotes. Assim, a ocorrência de erros leva a um aumento das tentativas de transmissão futura, o que tende a produzir um efeito de memória nas perdas. É justamente este efeito que se procura demonstrar nesta dissertação.

O modelo clássico de Gilbert-Elliott, utilizando cadeia de Markov com dois estados é utilizado em diversos estudos, como em (KRUNZ; KIM, 2001). Neste estudo é analisado o atraso de pacotes em uma rede sem fio, sendo que para a medição das rajadas de tráfego o processo de chegada do tráfego foi modelado como um processo de dois estados (*on-off*) contínuo e para a medição do canal foi utilizada a cadeia de Markov de dois estados (BOM e RUIM), onde cada estado é associado à uma determinada taxa de erro de bit (*bit error rate* - BER). Assim, este estudo utiliza o modelo de Gilbert – Elliott com distribuição exponencial de probabilidade para o

estado BOM e RUIM, mas com diferentes valores médios, de 0,1 e 0,0333 segundos, respectivamente.

No estudo de (LEE; CHANSON, 2002) são utilizadas duas cadeias de Markov em uma rede sem fio, sendo uma cadeia de Markov para calcular a probabilidade de perda de pacotes e outra para determinar a distribuição do atraso dos pacotes. Os pacotes são transmitidos através de um canal com erro, modelado por uma cadeia de Markov de dois estados. Se a transmissão falhar, o pacote é retransmitido até que um limite de atraso seja atingido. Caso excedido o limite de tempo, o pacote é descartado e começa a transmissão do próximo pacote. Este processo de descarte tem um impacto significativo sobre a probabilidade de perda de pacotes, mas raramente é considerado em outros modelos de Markov. Neste estudo os resultados obtidos demostram que a probabilidade de perda de pacotes é significativamente afetada pelo limite do tempo de atraso e pela probabilidade do canal permanecer no estado de falha, sendo praticamente independente da taxa de chegada.

O modelo simplificado de Gilbert é aplicado em (VELLA; ZAMMIT, 2013) para analisar a perda de pacotes na transmissão de multicast em uma rede 802.11n. No estudo são utilizadas múltiplas antenas e uma diversidade de posicionamento, visando reduzir a taxa de erro durante a transmissão, pois o mecanismo de envio utilizado não possui a confirmação de recebimento. O ambiente de teste para levantamento de dados foi realizado em três configurações distintas, sendo analisada a correlação espacial entre os receptores, o tamanho das rajadas de erro, o valor médio das rajadas de tráfego recebido sem erros e os parâmetros do modelo simplificado de Gilbert. Quanto à aplicação do modelo simplificado, o estudo demonstra que este modelo apresenta uma boa aproximação para o valor médio da taxa de perdas percebidas pelos dispositivos utilizados no teste. Como conclusão é ressaltada a influência da correlação espacial, afetando principalmente os dispositivos mais distantes.

2.2 Outros modelos de perdas de pacotes em redes WiFi

Um estudo que utiliza o modelo estendido de Gilbert, aplicando cadeias de Markov para modelar a transmissão de pacotes em redes wireless, é desenvolvido por (YOUNESIAN et al., 2014). Este estudo desenvolve um modelo para determinar a taxa de perda de pacotes, apresentando o resultado da simulação do modelo desenvolvido quando aplicado em dispositivos com interfaces 6LoWPAN (*IPv6 over Low power Wireless Personal Area Networks*) e WiFi. O modelo proposto no estudo é baseado nas duas categorias do modelo estendido de Gilbert, que são o *Reception Run-Lengths* (RRL) e o *Loss Run-Lengths* LRL. O modelo RRL de ordem m é composto de $m + 1$ estados: $\{S_0, S_1, \dots, S_m\}$, sendo que o sistema evolui para o próximo estado, a partir de S_0, a cada pacote recebido, e retorna ao estado S_0 caso ocorra a perda de pacote. O modelo LRL é semelhante ao RRL, porém com o avanço dos estados a cada pacote perdido, retornando ao início quando um pacote é recebido. O resultado da simulação deste estudo identifica a probabilidade de permanência em cada um dos estados do RRL ou LRL, calculando a matriz de probabilidade de transição e a probabilidade de permanência no estado de perda. A simulação foi realizada considerando três diferentes distribuições de probabilidade para representar a condição do canal: constante, gaussiana e exponencial. E como resultado final do trabalho é demonstrado que a taxa de perda de pacotes pode ser reduzida em dispositivos utilizando múltiplas interfaces sem fio, quando comparado com dispositivos com apenas uma interface sem fio.

Outros estudos realizados sobre perdas de pacotes também apontam para a necessidade de um modelo mais complexo do que o de Gilbert-Eliott. Um destes estudos, baseado no tráfego de aplicações multimídia em redes WiFi, analisou o comportamento das rajadas de erros em diversos cenários de transmissão (ANGEJA; NAVARRO, 2005). Para cada cenário foi determinado o parâmetro da distribuição em função dos diferentes tamanhos de pacotes, característicos do tráfego de voz ou de vídeo. Neste estudo o modelo proposto utiliza duas séries de distribuições logarítmicas, sendo uma para a rajada de perda de pacotes e outra para o número consecutivos de pacotes recebidos. Um dos fenômenos observado neste estudo foi a característica de cauda pesada na distribuição estatística do número consecutivo de pacotes recebidos, bem como a existência de rajadas de perdas.

O estudo sobre o tráfego de vídeo em uma rede WiFi realizado por (RUSS; HAGHANI, 2009) também conclui que a ocorrência de erros em redes WiFi não pode ser representada apenas pelo modelo de Gilbert-Elliott. Porém neste, o foco foi a análise das possíveis causas para a ocorrência de erros, tais como os dispositivos de rede utilizados, o método de adaptação de velocidade do padrão 802.11g ou interferências externas, indicando a necessidade de um estudo mais aprofundado, não realizando uma modelagem dos dados obtidos. A conclusão deste trabalho apenas apontava para que o modelo de distribuição dos surtos de erros seria uma combinação do modelo clássico de Gilbert-Elliott com uma distribuição de cauda pesada.

No estudo de (VIEIRA CARDOSO; REZENDE, DE, 2009) é proposto um modelo diferente para perda de pacotes, porém neste caso utilizando uma cadeia de Markov escondida (HMM - *Hidden Markov Model*), com duas estruturas, uma para a transição entre qualquer par de estados, chamado de estrutura geral, e outra apenas para estados adjacentes, chamado de *birth-death*. A utilização da estrutura com o *birth-death*, chamada de HMM3bd, apresentou um resultado muito melhor do que da estrutura geral, chamada de HMM3g, que é explicada pela natureza do processo de perdas cuja variação acontece rapidamente mas não abruptamente, ou seja, a qualidade do canal não varia instantaneamente. Foi então incrementada a quantidade de estados, observando-se uma melhora na precisão do modelo. Porém o valor ideal da quantidade de estados depende do conjunto de dados analisados. Este trabalho indica que o modelo HMM usando a estrutura *birth-death* é mais eficiente, até um limite de11 estados.

2.3 Distribuições de cauda pesada

Uma distribuição de probabilidade possui cauda pesada se a distribuição complementar possuir o seguinte comportamento:

$$Pr\{X > x\} \sim c.x^{-\alpha}, \qquad x \to \infty \tag{3}$$

onde α é o parâmetro de forma e c é uma constante positiva.

Uma das distribuições típicas de cauda pesada é a distribuição de Pareto, cuja distribuição acumulada de probabilidade é dada por:

$$Pr\{X \leq x\} = 1 - \left(\frac{\beta}{x}\right)^{\alpha}, \qquad x > 0, \qquad \alpha > 0 \tag{4}$$

onde α é o parâmetro de forma e β é o parâmetro de localização, cujos valores determinam a forma da curva da distribuição. Para valores de α menores que 1 a média e a variância não convergem e para valores de α entre 1 e 2 a distribuição apresenta variância não convergente.

Um dos métodos mais utilizados para testar a aderência à distribuição de cauda pesada é o gráfico da distribuição complementar, dada por $\bar{F}(x) = 1 - F(X)$ com $F(x) = Pr\{X \leq x\}$, em escala logarítmica, chamado de *Log-Log Complementary Distribution* (LLCD). O LLCD pode ser utilizado para determinação do parâmetro α buscando-se uma invariância dada por:

$$\frac{d\, log(\bar{F}(x))}{d\, log(x)} = -\alpha \tag{5}$$

cujo coeficiente angular é uma estimativa de α.

2.4 Séries Temporais

Sendo X_t uma série temporal discreta, denotada por X_t, $t = 0, 1, \dots, N$, onde *t* representa uma amostragem periódica ou uma série de intervalos de comprimento fixo. Diz-se que X_t é estritamente estacionária se $\{X_{t1}, X_{t2}, \dots, X_{tn}\}$ e $\{X_{t(1+k)}, X_{t(2+k)}, \dots, X_{t(n+k)}\}$ possuem a mesma distribuição conjunta para todo *n*. Uma série estacionaria não apresenta dependência temporal em função de seus valores passados. Pode-se caracterizar a dependência entre os valores da série em diferentes intervalos de tempo através da avaliação da função de autocorrelação (ACF - *autocorrelation function*), denotada por $\rho(k)$.

$$\rho(k) = \frac{E.[(X_t - \mu).(X_{t+k} - \mu)]}{\sigma^2} \tag{6}$$

onde μ é a média, *k* é o deslocamento e σ é o desvio padrão de X_t.

A ACF mede a similaridade entre uma série de X_t e uma versão deslocada da própria série X_{t+k}, com $-1 \leq \rho(k) \leq 1$. Quando $\rho(k)$ for zero, não há auto correlação

com deslocamento *k*. Se $\rho(k)$ for 1 ou -1 o valor presente está relacionado, respectivamente, de modo direto ou inverso com o valor de *k* passado.

2.5 Dependência temporal de longa duração

A estacionariedade de segunda ordem ocorre quando a ACF é semelhante, independentemente do intervalo considerado. Ou seja, a ACF de duas amostras não sobrepostas deslocadas no tempo deve ser equivalente.

Algumas características que vem sendo identificadas na análise de tráfego da Internet durante as últimas décadas são a auto similaridade, distribuições de cauda pesada e dependência temporal de longa duração, conhecida como *Long-Range Dependence* (LRD). Apesar de seu uso já estar bastante difundido, a análise da LRD é prejudicada pela dificuldade em realmente identificar a dependência de longa duração e estimar seus parâmetros de forma precisa. A LRD significa que o comportamento de um processo com dependência temporal apresenta auto correlações significativas durante uma escala de tempo muito ampla.

Uma série estacionária X_t tem dependência de longa duração se sua autocorrelação decai à zero tão lentamente que a sua soma não converge, ou seja, se $\sum_{k=1}^{\infty}|\rho(k)| = \infty$. Intuitivamente, uma série do tipo LRD possui memória porque a dependência entre valores mesmo muito distantes é significativa, mesmo com grandes deslocamentos de tempo (KARAGIANNIS et al., 2004).

Uma série temporal X_t é auto similar se:

$$X_t \overset{\cong}{d} a^H X_t, a > 0 \tag{7}$$

onde a igualdade refere-se a igualdade em termos de distribuição de probabilidade, a é um fator de escala e *H* é o parâmetro de Hurst.

Outra definição de auto similaridade é dada em função da autocorrelação da série (PARK; WILLINGER, 2000), que deve seguir a expressão:

$$\rho(k) = \frac{1}{2}[(k+1)^{2H} - 2k^{2H} + (k-1)^{2H}], \qquad 0{,}5 < H < 1 \tag{8}$$

Séries auto similares normalmente possuem estacionariedade de segunda ordem. Por outro lado, auto correlações com decaimento rápido caracterizam dependência de curta duração. Uma série temporal com $0{,}5 < H < 1$ tem dependência

de longa duração (LRD), e quanto mais próximo de 1 for o valor do parâmetro de Hurst, maior será a dependência temporal (KARAGIANNIS et al., 2004).

2.6 Modelo ARIMA

ARIMA é uma técnica de modelagem linear. Nesta técnica, primeiro é verificada a estacionariedade dos dados da série temporal. Se eles não são estacionários, é realizada uma operação diferencial. Enquanto os dados não forem estacionários, esta diferenciação é realizada *d* vezes, até atingir a estacionariedade. Assim, o valor de *d* necessário para que a estacionariedade seja atingida é chamado de ordem de integração do método ARIMA (BABU; REDDY, 2014).

Para realizar a diferenciação da série, buscando atingir a estacionariedade, o modelo ARIMA utiliza o processo de diferenciação da série, baseada no operador Nabla, que é definido por $\nabla^d X_t = X_t - X_{t-d}$, também chamado de operador diferencial. Assim, o operador Nabla (∇) atua como um operador de deslocamento. Porém, para tornar uma determinada série estacionária, pode ser necessário repetir o processo de diferenciação diversas vezes. Deste modo, o operador Nabla (∇) pode representar diversas etapas de diferenciação.

Após a aplicação do operador Nabla, a série deve ser estacionária. Na sequência, a série será modificada retirando-se a média $\tilde{X}_t = X_t - \mu$.

$$\tilde{X}_t = \emptyset_1 . \tilde{X}_{t-1} + \emptyset_2 . \tilde{X}_{t-2} + \cdots + \emptyset_p . \tilde{X}_{t-p} + a_t - \theta_1 . a_{t-1} - \cdots - \theta_q . a_{t-q} \quad (9)$$

onde os termos $\emptyset_1$ até $\emptyset_q$ são os coeficientes do modelo auto regressivo (AR - *Autoregressive*) e os termos θ_1 até θ_p são os coeficientes do modelo de média móvel (MA – *Moving Average*). O modelo de série temporal é denotado por ARIMA (*p*, *d*, *q*). O modelo ARMA pressupõe que a_t é um ruído branco, apresentando distribuição gaussiana com média zero, sendo a variância de a_t um parâmetro do modelo. O procedimento de modelagem ARIMA tem três etapas: diferenciação inicial, identificando-se o valor do operador nabla, identificação de *p* e *q*; e estimação dos coeficientes do modelo.

A identificação de *p* e *q* é feita usando-se as funções de auto correlação e de auto correlação parcial (BOX; JENKINS, 1990). Os coeficientes podem ser estimados usando-se o método de Box-Jenkins. Das várias abordagens possíveis, as

abordagens de estimativa Gaussiana de máxima verossimilhança, chamadas de *Gaussian Maximum Likelihood Estimation* (GMLE), são geralmente utilizadas para a estimativa dos parâmetros modelo ARIMA (YAO; BROCKWELL, 2006). O modelo é validado normalmente analisando-se o ruído, que deve possuir distribuição normal e não deve estar auto correlacionado no tempo. Depois que todos os coeficientes do modelo foram calculados, os valores de série temporal podem ser previstos utilizando-se os valores anteriores e os coeficientes do modelo. Os modelos ARIMA podem prever os dados de uma série temporal linear com muito boa precisão (BABU; REDDY, 2014).

2.7 Modelo FARIMA

No modelo *Fractional Auto Regressive Integrated Moving Average* (FARIMA) a ordem de integração *d* possui um valor no intervalo $0 \leq d \leq 1$, ou seja, será um valor fracionário. Este modelo apresenta dependência de longa duração (LRD), e o parâmetro de Hurst está relacionado com o parâmetro *d* através da relação $H = d + \frac{1}{2}$.

O operador de diferenças é redefinido como:

$$\nabla^d X_t = \sum_{i=0}^{d} \binom{d}{i} (-1)^i X_{t-i}, -\frac{1}{2} < d < \frac{1}{2} \tag{10}$$

E o coeficiente binomial pode ser interpretado como:

$$\binom{d}{i} (-1) = \frac{\Gamma(-d+i)}{\Gamma(-d)\Gamma(i+1)} \tag{11}$$

onde Γ é a função gama, que é uma extensão da função fatorial com o argumento deslocado de uma unidade. Para números complexos com parte real positiva, é definida pela integral: $\Gamma(\mathrm{z}) = \int_0^\infty x^{z-1}\, e^{-x} dx$ (MEYER, 1983).

2.8 Clusterização

O uso da metodologia de identificação dos estados B e R proposto pelo modelo de Gilbert-Elliott não pode ser aplicado em nosso caso pois o objetivo é contestar os resultados deste modelo. Desta forma buscou-se uma alternativa para a identificação dos estados B e R utilizando-se métodos de agrupamento ou clusterização. Foram então pesquisados os métodos possíveis de serem aplicados para análise e agrupamento dos dados obtidos a partir das amostragens realizadas.

No processo de análise de dados, existem quatro métodos diferentes que podem ser aplicados na aprendizagem (WITTEN et al., 2011):

- Classificação: a partir de um conjunto de dados já classificados espera-se obter a regra de classificação para os dados desconhecidos.
- Associação: busca-se uma característica qualquer dentro de uma associação, e não apenas características de uma classe em particular.
- Clusterização: busca-se formar grupos de dados.
- Previsão numérica: o resultado a ser previsto não é uma classe discreta, mas uma quantidade numérica.

No processo de clusterização os elementos poderão pertencer apenas a um grupo, de maneira exclusiva. Porém, poderão ocorrer superposições, onde um elemento pode pertencer a grupos diferentes, de maneira estatística, onde pertence a um grupo com uma certa probabilidade, ou hierárquica, onde um nível mais alto de agrupamento vai definir os grupos de maneira mais refinada. Assim, a natureza dos mecanismos envolvidos no fenômeno estudado é que irá definir qual o processo será mais adequado. Porém, como estes mecanismos nem sempre são bem conhecidos, a escolha acaba sendo definida em função das ferramentas disponíveis para a clusterização.

Portanto, o sucesso do processo de clusterização visa a obtenção de conjuntos com alto grau de similaridade, quando os elementos dentro do conjunto são muito semelhantes. Porém, os conjuntos/clusters devem ser bastante diferentes entre eles, ou seja, os elementos devem ser bastante heterogêneos em relação aos elementos dos outros clusters.

O processo de clusterização tem então três parâmetros envolvidos:

- Similaridade: comparação entre o número de atributos que dois elementos têm em comum, em relação ao total de atributos que existem entre eles. Objetos que tem todos os atributos iguais possuem similaridade igual a 1 e quando não tem nada em comum terão similaridade igual a zero.
- Dissimilaridade: é o complemento da similaridade, caracterizando o número de atributos diferentes entre dois elementos em relação ao número total de tributos definidos.
- Distância: é um conceito geométrico da proximidade entre os objetos, baseado na medição dos atributos. Assim, a definição de índices e métricas adequados para este parâmetro é de fundamental importância na análise dos clusters.

Um bom método de agrupamento deverá fornecer grupos que possuam alta similaridade intragrupo e baixa similaridade intergrupo. Desta forma a qualidade do resultado de um agrupamento depende tanto da medida de similaridade usada pelo método como da sua implementação. A qualidade de um método de agrupamento pode também ser medido pela sua habilidade para descobrir os padrões escondidos.

Para o processamento dos dados deve ser construída uma matriz de Dados e uma matriz de Dissimilaridade, que é definida a partir da matriz de Dados pela aplicação de um método de medição da Dissimilaridade.

A Proximidade é uma função que mede a similaridade ou a dissimilaridade entre um par de elementos, sendo que deve ser utilizada uma função a parte para a medição da qualidade de um grupo. As funções de proximidade dependem da escala das variáveis: proporcional, intervalar, ordinal, nominal, binária ou mista, porém, a definição do grau de similaridade é quase sempre resultado de uma avaliação subjetiva.

Medição de Similaridade: A proximidade entre dois dados x_i e x_j é representada por $d(x_i, x_j)$, que deve ser determinada por uma métrica, sendo a mais utilizada a chamada distância de Minkowski (SINGH et al., 2013), dada por:

$$d(x_i, x_j) = \sqrt[p]{\sum_{k=1}^{d} (|x_{ik} - x_{jk}|)^p}, p \geq 1 \quad (12)$$

onde *d* é a distância entre os dados e *p* é a dimensão da representação dos dados.

Com $p = 1$ temos a chamada Distância de Manhattan, que é representada por:

$$d(x_i, x_j) = \sum_{k=1}^{d} (|x_{ik} - x_{jk}|) \quad (13)$$

Quando $p = 2$, temos a chamada distância euclidiana, que é a métrica mais utilizada quando os elementos possuem valores contínuos, avaliando a proximidade de dados representados em duas dimensões, dada por:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (|x_{ik} - x_{jk}|)^2} \quad (14)$$

Uma das desvantagens do uso da Distância Euclidiana é a possibilidade da dominância de um atributo sobre os demais, o que pode ser resolvido com a normalização dos valores.

Os algoritmos de clusterização são classificados em:

- Hierárquicos Aglomerativos
- Hierárquicos Divisivos
- Cluster Fixo

Nos algoritmos hierárquicos a construção dos clusters segue uma estrutura hierárquica, agrupando os elementos em um número de clusters que variam de um cluster até a quantidade total de elementos. Os algoritmos de Cluster Fixo, também chamados de métodos de particionamento, criam os agrupamentos de acordo com o número de clusters desejado, que deve ser definido inicialmente para a execução do algoritmo.

Na análise aglomerativa, o processo inicial consiste em colocar os elementos individualmente e os dois elementos que tenham um maior grau de similaridade serão agrupados, formando um cluster. O processo de agrupamento continua, analisando os demais elementos e procurando a maior similaridade entre todos eles, de forma a obter os grupos que sejam o mais similares possíveis.

2.9 O método k-means

Entre os algoritmos do método hierárquico aglomerativo, uma das técnicas clássicas de clusterização é a chamada k-means. O processo inicia-se com a definição do número de clusters que se deseja obter, que será o parâmetro *k*. Então são escolhidos aleatoriamente *k* elementos como centro dos clusters, sendo associados todos os demais elementos ao seu cluster mais próximo, de acordo com a métrica definida pela distância Euclidiana, mostrada na Equação 15. Após a formação dos clusters, é calculado o centroide de cada cluster, obtido pelo cálculo da média dos elementos. Estes centroides então serão definidos como o centro dos seus respectivos clusters, e o processo é repetido, para estes novos centros, associando-se os elementos ao centro mais próximo, modificando-se a formação dos clusters. O processo é repetido até que os centros dos clusters não sejam mais alterados, formando-se então os *k* clusters definitivos.

Este método é simples e efetivo, e este algoritmo pode ser escrito para lidar eficientemente com conjuntos de dados muito grandes, por isso pode ser útil em casos onde outros métodos falharem. Porém o resultado final é bastante dependente da escolha inicial dos centros dos clusters. Deste modo, é necessário repetir o processo de clusterização diversas vezes, com centros iniciais diferentes, comparando-se o resultado para determinar qual foi o melhor conjunto final, que é aquele que apresenta menor distância quadrática total. Também, soluções como número de clusters finais diferentes podem apresentar estruturas totalmente diferentes, o que torna este método ineficiente se não for conhecido inicialmente a quantidade ideal de clusters.

Para examinar os resultados a principal ferramenta gráfica utilizada é conhecida como dendograma, que é uma exibição no formato de árvore que lista os objetos que são agrupados, ao longo do eixo *x*, e a distância em que o cluster foi formado, ao longo do eixo *y*. As distâncias ao longo do eixo *x* não tem sentido em um dendograma pois as observações são igualmente espaçadas para facilitar a leitura.

Se escolhido um valor no eixo vertical e percorrido o dendograma horizontalmente, contando o número de linhas cortadas, obtém-se a quantidade de clusters formados, pois cada linha representa um grupo de elementos que foram agrupados em cada cluster. Para identificar os elementos de cada grupo é necessário percorrer o dendograma até a sua base, onde estão identificados os elementos.

Como o eixo *y* representa quão próximos estavam os elementos quando eles foram fundidos em clusters, os clusters cujos ramos estão muito próximos entre si provavelmente não são muito confiáveis. Mas se há uma diferença grande entre o último cluster mesclado e o atualmente mesclado, isso indica que os clusters formados provavelmente estão bem estruturados, conforme a estrutura dos dados mostrada ao longo do eixo *y*.

2.10 Clusterização k-means no software R

O método de clusterização k-means está disponível no software R através da função `kmeans`. O processo começa escolhendo os elementos de *k* para servir como centros dos clusters. Então, a distância de cada um dos demais elementos para cada um dos centros dos clusters *k* é calculada e os elementos são colocados no cluster ao qual eles estão mais próximos. Após cada elemento ser colocado em um cluster, o centro dos clusters é recalculado, e cada elemento é verificado para ver se poderia estar mais próximo de outro cluster. O processo continua até que não existam mais elementos para serem movidos para outros clusters.

A biblioteca *cluster* do software R fornece uma alternativa moderna para algoritmos k-means de clusterização, conhecida como *pam* (*Partitioning around Medoids*). O termo *medoid* refere-se a identificação de um ponto dentro de um cluster, de forma que a soma das distâncias entre ele e todos os outros membros do cluster seja mínima. O algoritmo *pam* requer que o número de clusters desejados seja conhecido, assim como o *k-means*, mas ele faz um processamento maior do que o *k-means* para assegurar que os *medoids* encontrados são verdadeiramente representativos dos elementos dentro de um determinado cluster.

No método *k-means* os centros dos clusters são somente recalculados após todos os elementos terem passado de um cluster para outro. No *pam* as somas das distâncias entre objetos dentro de um cluster são recalculadas constantemente, à medida que os elementos são movidos, o que deve fornecer uma solução mais confiável.

Além disso, como um subproduto da operação de cluster, ele identifica as observações que representam os *medoids*, e estas observações (uma por cluster) podem ser consideradas um exemplo representativo dos membros desse cluster, que

pode ser útil em algumas situações. O algoritmo *pam* exige que a matriz distância seja totalmente calculada para facilitar o recálculo dos valores de *medoids* exigindo consideravelmente mais recursos de computação do que o *k-means*. Como no *k-means*, não há nenhuma garantia de que a estrutura obtida com um pequeno número de clusters seja mantida quando for aumentado o número de clusters.

Para os métodos de clusterização hierárquica, o dendograma é a principal ferramenta gráfica para analisar uma solução de cluster. Quando utilizadas as funções `hclust` ou `agnes` para realizar uma análise de cluster pode-se obter o dendograma utilizando a função `plot` sobre o resultado da clusterização.

A primeira etapa no método hierárquico é o cálculo da matriz de distância, sendo que para um conjunto de dados com *n* observações, a matriz de distância terá *n* linhas e *n* colunas. Cada elemento (*i*, *j*) da matriz de distância será a diferença entre o elemento i e o elemento j. Existem duas funções que podem ser usadas para calcular as matrizes de distância no software R: a função `dist`, que está incluída em cada versão do R, e a função `daisy` que é parte da biblioteca `cluster` e oferece alguns recursos que não estão na função `dist`. Cada função oferece uma escolha de métricas de distância que propiciam diferentes análises sobre a estrutura dos dados.

O método padrão do `hclust` para o cálculo da matriz de distância é o acoplamento completo. Neste método, quando um cluster é formado, sua distância a outros elementos é calculada como a distância máxima entre quaisquer elementos dentro do cluster.

3 COLETA E TRATAMENTO DE DADOS

Neste capítulo são apresentados os materiais utilizados, bem como o procedimento de coleta e processamento de dados.

3.1 Materiais e métodos

O estudo contemplou duas fases distintas quanto à necessidade de recursos de hardware e software. Na fase de coleta de dados foram utilizados dois computadores para transmissão e recepção do tráfego de teste, e os softwares para o envio e coleta dos dados. Na fase de análise foi utilizado um computador com software para análise estatística, além de programas desenvolvidos em linguagem C para o tratamento dos dados.

Para a fase de coleta de dados a metodologia adotada considerou a realização da medição em datas e horários distintos, com diversos níveis de utilização e quantidade de usuários na rede amostrada.

3.2 Procedimentos de teste

Para a coleta de dados foram utilizados dois computadores, com sistema operacional Linux, um deles operando como cliente e outro como servidor, conforme mostrado na Figura 2.

O computador cliente se conectava à rede através de um Access Point, no modelo infra estrutura, executado um programa que fazia o envio de pacotes de controle para um servidor, que armazenava os dados recebidos. A conexão do servidor foi feita através da rede cabeada, conectando-se através de um switch conectado ao AP.

Em uma primeira fase de testes, para a geração e armazenamento do tráfego de teste foram utilizados os softwares iPerf, para o envio de pacotes, e o Tcpdump, para recebimento e registro dos pacotes. O Tcpdump é um software com distribuição livre, que possibilita monitorar os pacotes recebidos ou transmitidos em uma determinada interface do computador, gravando uma descrição do conteúdo dos

pacotes, com a informação de tempo contendo os valores de hora, minutos, segundos e frações de segundo. O iPerf é uma ferramenta para medições ativas em redes IP, possibilitando o ajuste de vários parâmetros relacionados a temporização, buffers e protocolos (TCP, UDP, SCTP com IPv4 e IPv6). Porém como o identificador de pacotes utilizado pelo Tcpdump é o identificador contido no cabeçalho do pacote IP, de 16 bits, a identificação de pacotes ficou limitada à 65.536 valores distintos. Com isto, nos casos em que a quantidade de pacotes perdidos era maior que 65.536, o identificador se repetia, não sendo possível o processamento correto dos dados na fase seguinte. A solução foi a elaboração de um programa em C, operando como um cliente UDP, com um identificador de pacotes de 64 bits, não se repetindo o valor do identificador durante todo o tempo de teste. Foi elaborado um programa em C para atuar como servidor e realizar o recebimento dos dados, sendo utilizados estes programas para a realização dos testes, e não mais os programas Tcpdump e iPerf. Ambos os programas foram desenvolvidos utilizando sockets em sistemas Linux.

O programa de teste executado no cliente possibilitava a configuração de taxa de transmissão e o tamanho do pacote do fluxo de controle, inserindo uma estampa de tempo e o número sequencial de cada pacote transmitido. Deste modo, com a informação de tempo seria possível avaliar o correto espaçamento ente eles e, se necessário, analisar a variação do tempo de chegada. O identificador sequencial dos pacotes do fluxo de teste foi o principal dado utilizado, pois a identificação dos estados de erro da rede foi baseada no não recebimento dos pacotes enviados, o que foi identificado pela análise deste dado. No servidor, o programa identificava os pacotes recebidos do cliente, gravando em arquivo o identificador de pacote, a estampa de tempo de envio e a de chegada no servidor, para o tratamento e análise dos dados na fase seguinte.

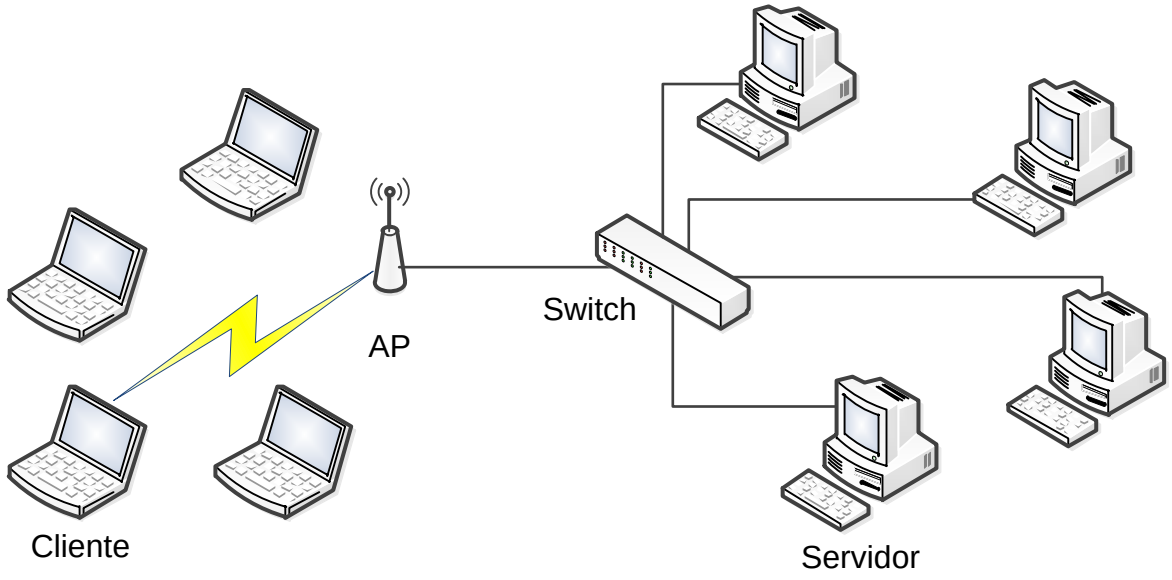


Figura 2 – Diagrama do ambiente de teste

A quantidade de dados a serem transmitidos pelo procedimento de teste da rede foi definida de forma a não interferir significativamente no nível de ocupação da rede, pois um tráfego de teste muito elevado poderia, inclusive, levar a um congestionamento. Os pacotes de controle foram definidos com tamanho de 100 bytes e enviados com uma taxa de 1000 pacotes por segundo. Somando-se o cabeçalho UDP (8 bytes), IP (20 bytes) e 802.11 (34 bytes), que totalizam 62 bytes, leva a um fluxo transmitindo um total de 162 bytes ou 1.296 bits por pacote. Deste modo o tráfego de teste gerado foi de 1,3 Mbps. Como o padrão da rede testada é o 802.11n, com taxa máxima observada na rede em teste de 144,45 Mbps, o tráfego de teste representa menos de 1% da capacidade da rede, com pouca influência no congestionamento da mesma. Com o envio de um pacote de teste a cada 1 milissegundo, de acordo com o Teorema de Nyquest, a taxa de amostragem de 1kHz representará o comportamento do sinal amostrado de até 500Hz (HAYKIN; VEEN, 1999). Assim o tráfego de controle poderá representar o estado da rede que tenha duração maior do que 2 milissegundos. A identificação dos estados da rede foi realizada baseada no tráfego enviado pelo cliente, não necessitando de nenhum mecanismo de confirmação de entrega, o que levou à utilização do protocolo UDP.

Além do envio e captura de pacotes, também foram registrados diversos parâmetros da rede, com a utilização do instrumento de medição AirCheck^{TM}Wi-Fi Tester, da Fluke Networks, colocado junto ao computador que transmitia o tráfego de teste, para realizar a medição dos parâmetros de rede vistos pelo cliente. O AirCheck

é um testador de redes WiFi, portátil, que permite verificar e solucionar problemas de redes 802.11 a/b/g/n/ac, verificando a disponibilidade, conectividade e segurança.

A amostragem da rede foi realizada com a coleta dos parâmetros da rede WiFi, medidos com o instrumento de teste AirCheck, incluindo a ocupação média, número de usuários no canal, canal utilizado, nível do sinal recebido, quantidade de APs no canal utilizado e tempo de duração do teste, e com o registro do tráfego de teste recebido pelo servidor. Os parâmetros de rede, duração e quantidade de pacotes transmitidos e recebidos são mostrados na Tabela 2.

Tabela 2 – Dados da rede e do tráfego de teste

Amostra	1	2	3	4
Data	05/04/17	06/04/17	18/04/17	14/06/17
Ocupação Média	65%	61%	66%	64%
Usuários no canal	19	10	41	39
Canal	1	1	11	11
Nível médio do Sinal	-59 dBm	-58 dBm	-39 dBm	- 64 dBm
APs no canal	7	8	8	9
Tempo de duração	1h	1h	45 min	1h
Pacotes transmitidos	3.312.992	3.304.135	3.257.448	3.311.930
Pacotes recebidos	3.225.218	3.172.692	2.888.677	1.917.873
Perdas	2,65%	3,98%	11,3%	42,1%

3.3 Dados das amostras

Para obter dados que representassem cenários com diversos níveis de utilização foram realizadas 4 amostras no ambiente de rede WiFi do Departamento de Engenharia Elétrica (DELT) da UFPR, em horários e locais distintos. Diariamente circulam nas dependências do DELT cerca de 1150 alunos e 52 professores, que contam com 6 Access Points para a conexão com a rede sem fio, com diversas redes distintas. Além das amostras apresentadas neste trabalho, foram feitas várias outras amostras, porém estas foram utilizadas apenas nas fases de desenvolvimento e validação dos programas e métodos de teste.

As amostras tiveram uma duração de uma hora, com diferentes quantidades de usuários, mas com um nível de ocupação bastante semelhante, entre 61% e 66%. Nas duas últimas amostras houve uma mudança de canal durante a realização do teste, identificada com a utilização do AirCheck, sendo que para a terceira amostra foram descartados os 15 primeiros minutos da amostra e processados os demais 45 minutos, eliminando-se o efeito de mudança de canal. Para a última amostra, foi realizado o teste por mais uma hora após a mudança de canal, para obter-se uma quantidade maior de amostras.

Os dados da rede, com a utilização do medidor da AirCheck, foram registrados a cada 5 minutos. Deste modo obteve-se uma base de dados da rede também com no mínimo 12 valores medidos, com a média dos intervalos de 5 minutos, para cada um dos parâmetros registrados, possibilitando a análise de subconjuntos das amostras maiores. No processo de análise foram utilizados estes subconjuntos, principalmente da primeira amostra que continha uma quantidade maior de dados, permitindo a comparação dos estados de rede associados com o nível de ocupação e número de usuários.

Além dos dados do nível de ocupação e da quantidade de usuários, também foram registrados outros parâmetros da rede, entre eles o nível do sinal, que é influenciado pela a distância entre o cliente e o AP, além dos demais obstáculos que possam existir entre eles. Nas duas primeiras amostras e na última amostra os equipamentos, computador de teste e AirCheck, estavam na mesma posição em relação ao AP ao qual estavam conectados, com o nível de sinal próximo à - 60dBm. A terceira amostra foi feita em uma distância menor, para coletarmos uma amostra também em um cenário com o usuário posicionado logo abaixo do AP, ou seja, na melhor condição possível, sem barreiras.

Comparando-se a quantidade de pacotes perdidos com a quantidade de usuários, constata-se que na primeira e segunda amostra houve uma redução para quase 50% do número de usuários, porém o percentual de perdas sofreu até um leve acréscimo. Este fato poderia estar associado ao fato do perfil de tráfego indicar o uso menos intenso da rede pelos usuários. Nas terceiras e quartas amostras a quantidade de usuários foi bastante semelhante, porém com um grande aumento do percentual de perdas, o que pode estar associado também ao perfil de tráfego dos usuários.

Em uma das amostras que foram realizadas, mas não incluída neste estudo, foi identificado um problema quando o cliente trocava de AP e de canal, o que levou ao descarte daqueles dados. Naquele cenário, o cliente estava recebendo o anúncio da mesma rede de dois APs operando no mesmo canal. Assim, em um determinado intervalo de tempo o tráfego de teste era enviado por um AP e em outro intervalo de tempo por outro. Com isso, nas trocas de AP os pacotes poderiam chegar ao servidor fora de ordem, gerando problemas na fase de processamento dos dados, e em alguns casos havia uma grande perda de dados no momento da troca de AP. Desta forma, para evitar este problema, durante todas as amostras incluídas neste trabalho foi acompanhado o estado da conexão ao AP, para garantir que não houvesse esta mudança.

3.4 Tratamento de dados

O processo do tratamento de dados foi realizado em diversas etapas, conforme ilustrado pela Figura 3, utilizando-se programas específicos desenvolvidos em linguagem C e o software R. O conjunto de dados coletados, pelo AirCheck e pelo fluxo de teste, estão identificados no diagrama da Figura 3 como Raw Data.

A primeira fase de processamento dos dados realizada foi a identificação dos pacotes não recebidos pelo servidor, completando a base de dados para a posterior clusterização. Para isto foi elaborado um programa em C que fazia a leitura sequencial do arquivo e inseria as linhas com os identificadores não recebidos, colocando o valor de tempo do último pacote enviado e com valor igual a zero no campo de tempo recebido. Nesta etapa foi gerada uma nova base de dados, contendo os pacotes recebidos e pacotes perdidos.

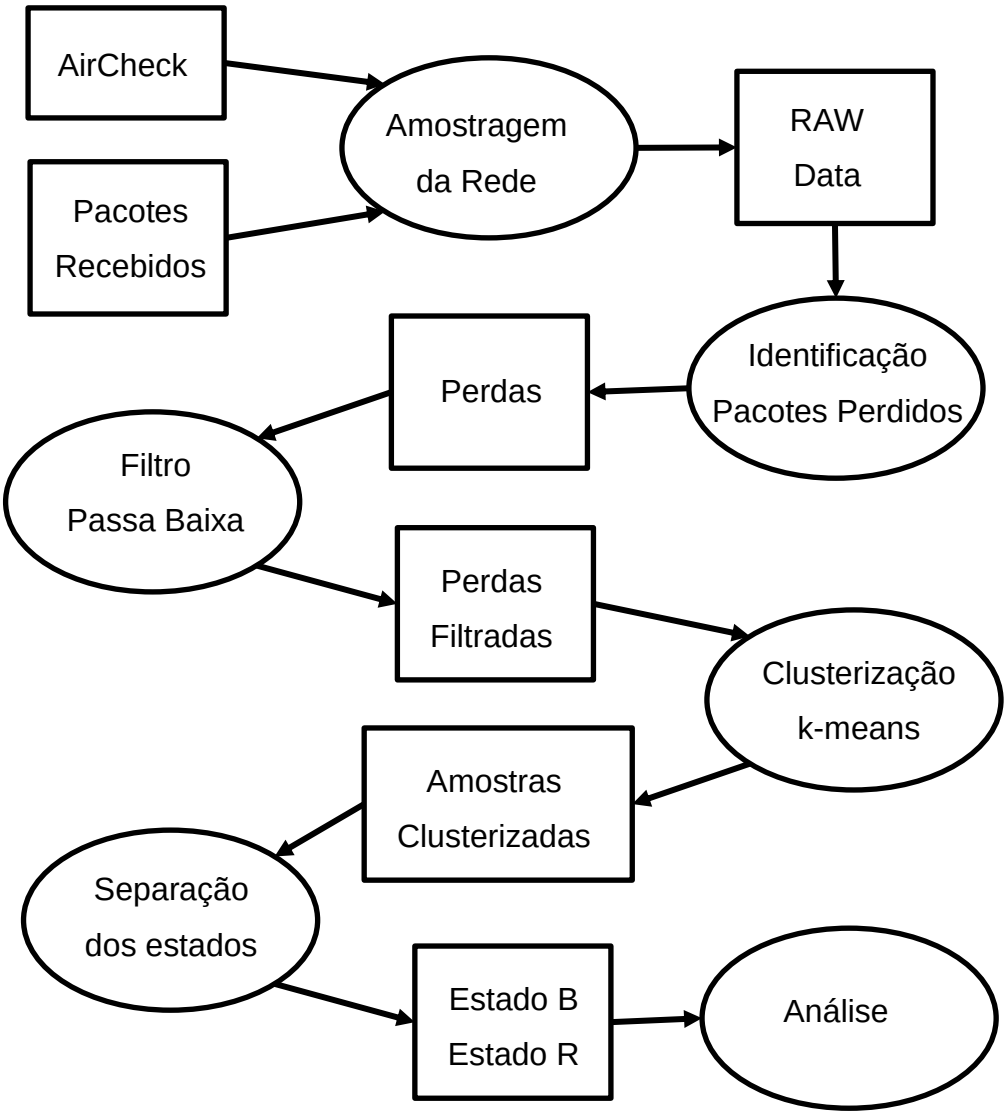


Figura 3 – Etapas do tratamento de dados

A sequência de pacotes de controle transmitidos foi transformada em uma série temporal, onde um pacote recebido foi representado com o número 1 e um pacote perdido foi representado como 0, gerando-se um arquivo para cada uma das amostras. Assim, o arquivo de dados de cada amostra continha cinco campos: tempo em que o pacote foi enviado, tempo em que o pacote foi recebido, identificador numérico sequencial e identificador de pacote recebido ou perdido.

A próxima fase foi a identificação da mudança de estado da rede, entre estado B e R, para a qual foi aplicado o método de clusterização k-means, descrito na seção 2.9, com a utilização do software R. Para a aplicação deste método é necessário definir-se um critério de parada do processo, pois ele irá gerar os clusters de maneira hierárquica, formando todas as combinações desde um único cluster até a quantidade

total de amostras. O valor definido foi igual à 2, pois deseja-se classificar os dados em uma formação com apenas dois clusters: estados B e R. O método k-means foi aplicado à série unidimensional, inicialmente com valores iguais à 0 e 1, representando a perda de pacotes.

Foi realizado o pré-processamento da série representando os pacotes perdidos/recebidos utilizando -se um filtro digital passa-baixa, definido como mostrado adiante. O objetivo da aplicação do filtro é definir um peso para os pacotes perdidos/recebidos de acordo com sua vizinhança, permitindo o funcionamento posterior do processo de clusterização. A função aplicada é mostrada na Equação 16.

$$X_t = X_{t-1}.\delta + (1-\delta).X_t \quad (16)$$

onde δ é o fator de amortização do filtro.

Para encontrar-se o valor ideal de δ, para o filtro passa-baixa, foi tomada uma amostra de 10.000 registros, aplicando-se diversos valores de δ e comparando-se os resultados do processo de clusterização. O primeiro parâmetro analisado foi o atraso entre o início de um estado R e a sua identificação no processo de clusterização, cuja variação é mostrada na Tabela 3. Este atraso aumentou com o valor de δ , de modo que estados R com duração menor do que 20 pacotes perdidos não eram nem reconhecidos quando aplicado um δ de 0,95. O outro parâmetro analisado foi o atraso no retorno ao estado B, após a clusterização, também mostrado na Tabela 3. Para o retorno ao estado B, os valores para δ 0,9 e 0,95 são os valores médios, pois este parâmetro também apresentava uma variação de acordo com a duração do estado R.

Tabela 3– Atraso na mudança de estado

Valore de δ	0,7	0,8	0,85	0,9	0,95
Entrada do estado R	2	2	4	5	20
Retorno ao estado B	1	3	4	7	8

Assim, o valor de δ igual a 0,7 foi o que representou menor atraso nas duas transições, sendo utilizado este valor para o filtro aplicado sobre a série. Desta forma, após a filtragem, a série representando os pacotes recebidos e perdidos foi transformada em um número real, variando entre 0 e 1.

Geradas as novas séries de dados para cada uma das amostras, com a aplicação do filtro, foi então aplicado o algoritmo k-means. O resultado do k-means

pode ser visualizado graficamente com a utilização do dendograma, mostrado na Figura 4, onde são agrupados hierarquicamente os dados a partir do peso 0, com um elemento em cada cluster, até o agrupamento total, com todos os elementos agrupados em um só cluster.

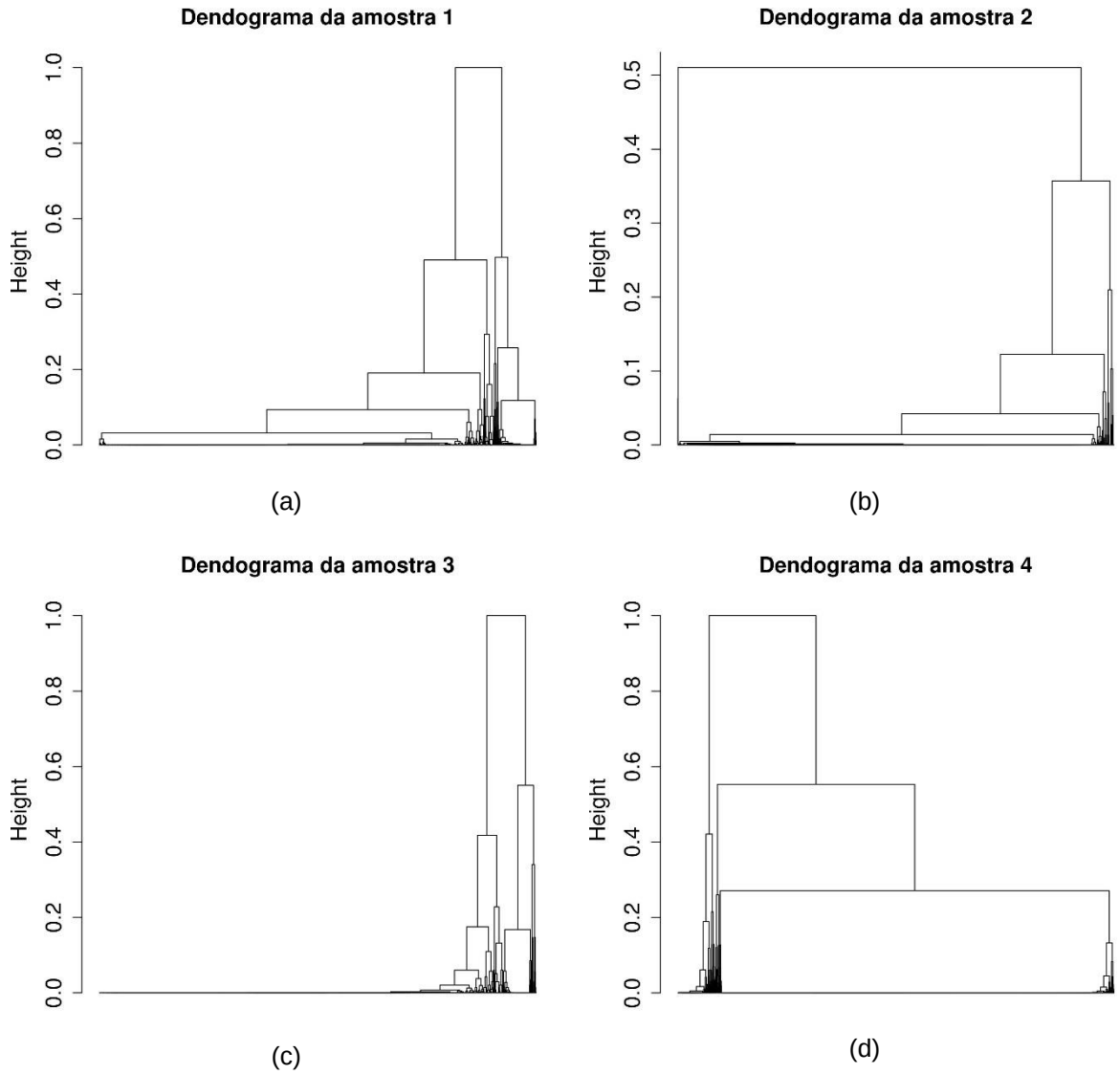


Figura 4 – Dendograma para as quatro amostras, com o peso dos agrupamentos formados representado ao longo do eixo *y*, e as amostras distribuídas ao longo do eixo *x*.

Como critério de formação dos clusters foi utilizado o valor de *k* igual à dois, pois o objetivo era a caracterização de dois estados distintos da rede: B e R. Como a formação de dois clusters representa mais de 40% do eixo y em três amostras e 30% na outra, isso indica que os clusters formados provavelmente estão bem estruturados. O resultado da clusterização foi a geração de um novo parâmetro, acrescentado às séries, representando o cluster ao qual o pacote pertencia.

4 RESULTADOS

Os resultados obtidos das medições realizadas, detalhando-se as características da rede para cada amostragem, são apresentados neste capítulo. Também é apresentada a modelagem dos dados, com os métodos discutidos anteriormente, indicando-se o nível de aderência aos dados amostrados.

4.1 Séries obtidas

Após a clusterização, foram geradas duas novas séries, contendo o tempo de duração dos estados B e R, que foram objeto de estudo, para identificação do modelo mais adequado para caracterizar o comportamento das perdas na rede WiFi. A quantidade de estados B e R de cada amostra, bem como os valores mínimos, máximos, médios, medianas e desvio padrão do tempo de duração dos estados, são mostrados nas Tabelas 4 e 5.

Tabela 4 – Dados do estado BOM das amostras

Amostra	1	2	3	4
Quantidade de Estados	5142	8809	9740	22583
Valor mínimo (s)	0,0010	0,0010	0,0010	0,0010
Valor máximo (s)	13,11	11,02	9,186	27,39
Média (s)	0,6309	0,3627	0,2418	0,082
Mediana (s)	0,1990	0,1270	0,0880	0,006
Desvio padrão (s)	1,1888	0,7118	0,5012	0,4609

Tabela 5 – Dados do estado RUIM das amostras

Amostra	1	2	3	4
Quantidade de Estados	5142	8809	9740	22583
Valor mínimo (s)	0,0010	0,0010	0,0010	0,0010
Valor máximo (s)	1,527	1,272	1,808	25,63
Média (s)	0,01343	0,01224	0,006356	0,06452
Mediana (s)	0,0010	0,0020	0,0010	0,017
Desvio padrão (s)	0,04377	0,03890	0,04307	0,3370

Os valores mínimos encontrados para todas a amostras, tanto do estado B quanto do estado R, de um milissegundo demonstra que, apesar de ter sido aplicado um filtro na fase de tratamento dos dados, ainda existem estados B e R com curta duração. Isto ocorre, pois, surtos muito curtos de mudança de estado acabam clusterizando apenas um pacote, pois quando iniciaria uma mudança de estado, a rede retorna para o estado anterior, permanecendo apenas o tempo equivalente a um pacote neste estado.

Quanto à duração dos estados R, poderia se esperar uma correlação do tempo de duração do estado com a quantidade de perdas, porém, o resultado é muito diferente disto. Da primeira para a segunda amostra o percentual de perdas quase dobrou, de 2,65% para 3,98%, porém o tempo médio de duração dos estados R permaneceu muito próximo. Para a terceira amostra o percentual de perdas aumentou ainda mais, para 3,98%, mas o tempo médio de duração dos estados R diminuiu para 50% do anterior, evidenciando que não existe uma relação entre o tempo médio do estado R e o percentual de perdas de pacotes. O que pode ser observado é que o crescimento apresentado no percentual de perdas de pacotes acompanhou o crescimento da quantidade de alteração entre os estados B e R, ou seja, ocorreram muito mais surtos de erros na rede, porém sem uma relação diretamente proporcional, pois a última amostra apresentou um aumento para 42,1%, quase quadruplicando o valor obtido na terceira amostra, mas a quantidade de mudanças de estados pouco mais que dobrou.

4.2 Verificação de correlação

A primeira análise realizada foi a verificação de uma eventual correlação estatística entre o tempo de duração dos estados B e R subsequentes, cujo resultado é mostrado na Tabela 6. Caso existisse uma correlação entre os estados B e R, os valores obtidos deveriam ser próximos a 1 ou -1, porém como os valores obtidos foram muito próximos a zero, isto indica que o tempo de duração dos estados B e R subsequentes não estão correlacionados.

Tabela 6 – Correlação entre o tempo de duração dos estados B e R subsequentes

Amostra	Correlação entre estados B e R
1	- 0,01153184
2	0,006807406
3	0,003391667
4	0,07600057

4.3 Verificação de dependência temporal

A próxima análise realizada foi a busca de uma possível dependência temporal na série formada pelo tempo de duração dos estados B e R, que não deveria existir no caso do modelo Gilbert-Elliott. Para identificar esta dependência temporal foi feita a análise da função de auto correlação para o tempo de duração dos estados.

A Figura 5 apresenta os gráficos das funções de auto correlação para as amostras 1, 2, 3 e 4. Nas quatro amostras pode ser identificada uma auto correlação com decaimento lento. A linha pontilhada identifica o intervalo de confiança de 95%. Assim, como o modelo de Gilbert-Elliott está baseado em cadeias de Markov, sem dependência temporal, para que o tempo de duração do estado B das amostras fosse representado por este modelo, os valores da ACF para deslocamentos maiores que 1 deveriam ficar próximos à zero, o que não ocorreu.

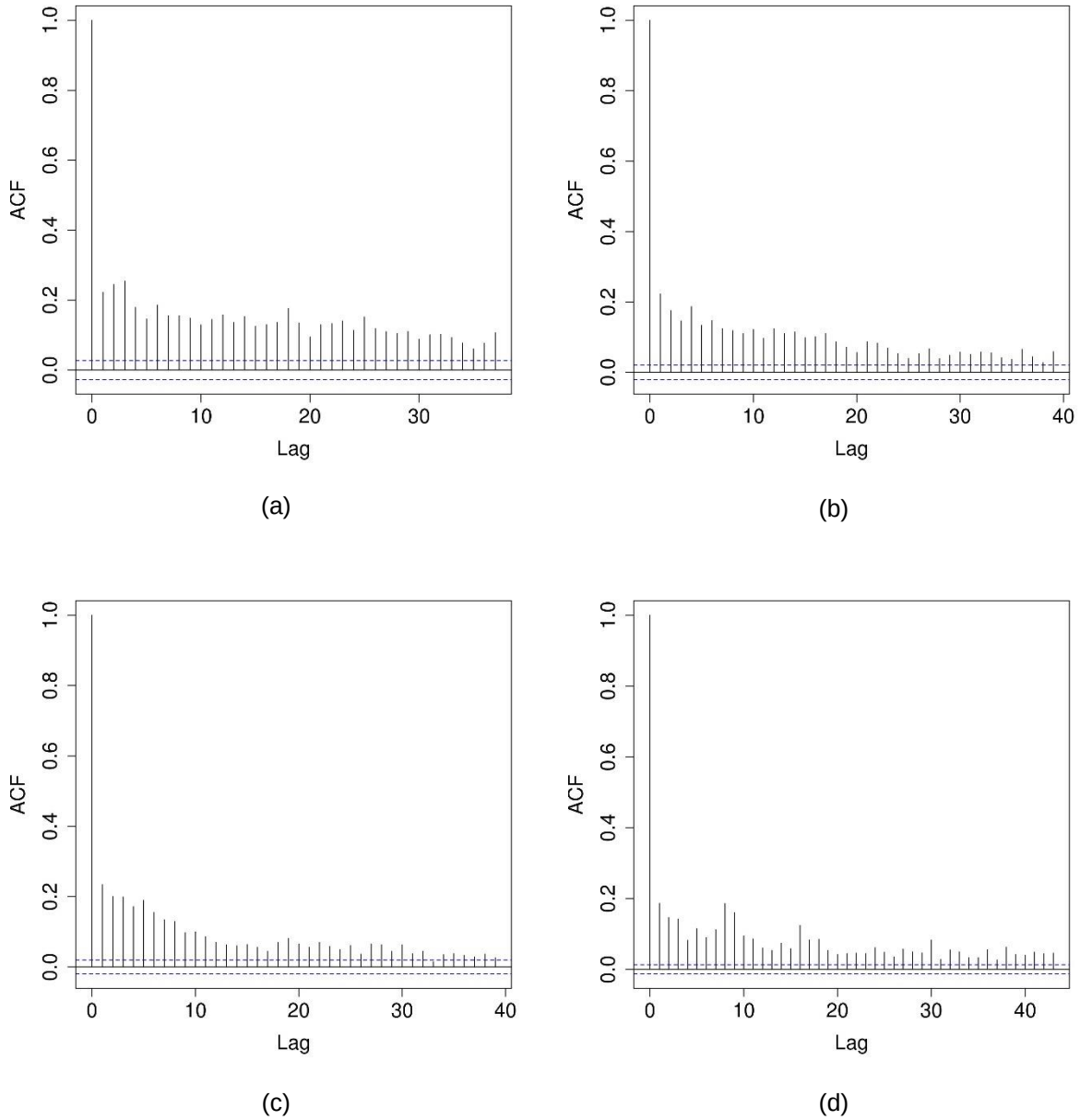


Figura 5 – ACF do tempo de duração dos estados B para a amostra 1 (a), amostra 2 (b), amostra 3 (c) e amostra 4 (d).

A função de auto correlação do tempo de duração do estado R das amostras é apresentado na Figura 6. Para o estado R também é identificada uma auto correlação significativa e com decaimento lento, o que indica que o tempo de duração do estado R das amostras também não segue o modelo de Gilbert-Elliott. Mesmo na última amostra, que não apresentou decaimento lento, identifica-se uma auto correlação até o deslocamento 6, indicando que também esta amostra não segue o modelo de Gilbert-Elliott.

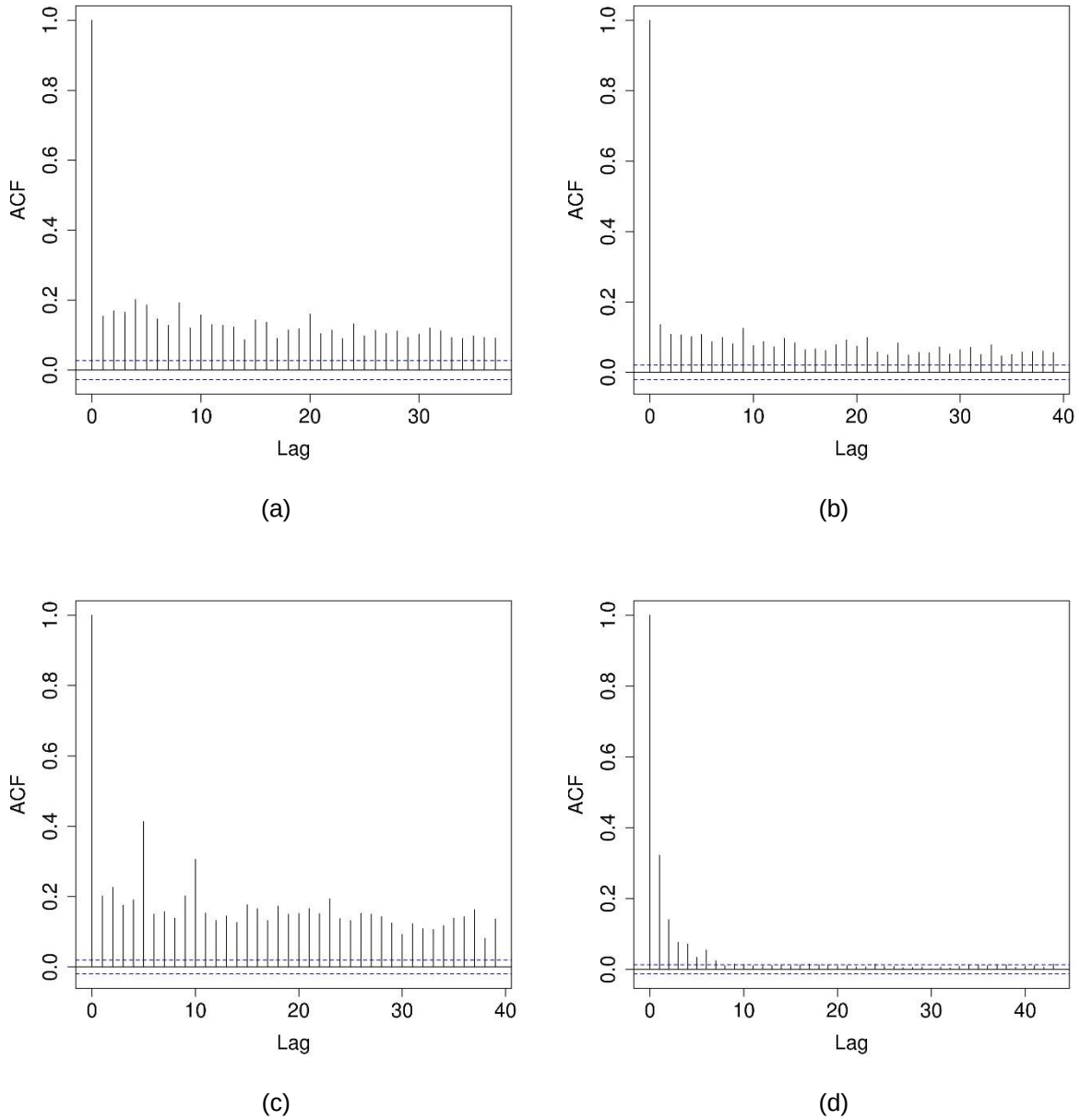
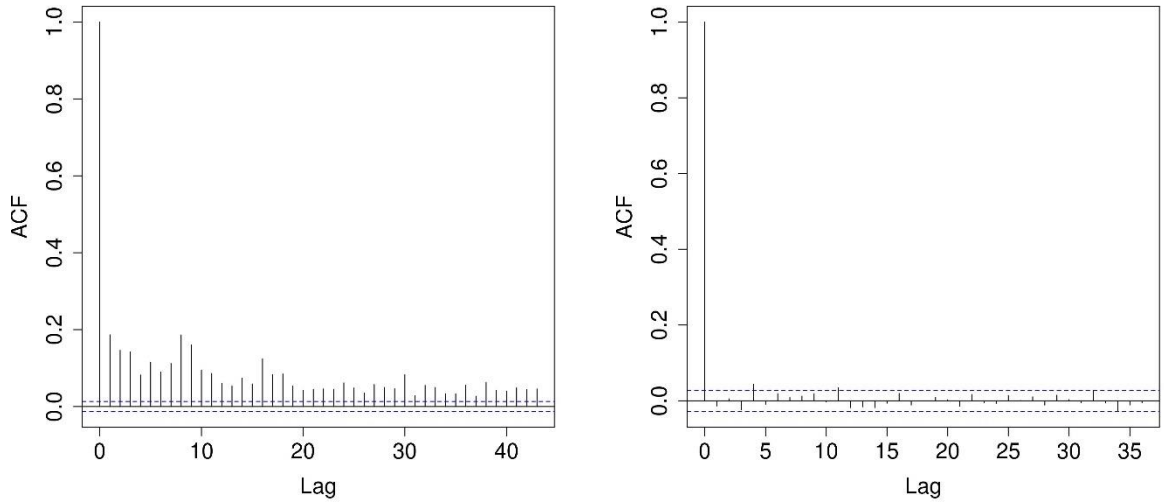


Figura 6 – ACF do tempo de duração dos estados R para a amostra 1 (a), amostra 2 (b), amostra 3 (c) e amostra 4 (d).

Também para efeito de comparação dos dados empíricos com o modelo de Gilbert-Elliott quanto à autocorrelação, foi elaborada uma simulação do modelo de Gilbert-Elliott parametrizada de acordo com os dados da amostra 4. A Figura 7(a) mostra a ACF dos dados empíricos e a Figura 7(b) mostra a ACF simulada. É possível observar que o modelo de Gilbert-Elliott não é capaz de reproduzir o comportamento dos dados empíricos.



(a) (b)

Figura 7 – ACF do tempo de duração dos estados B para a amostra 4 (a) e para uma simulação do modelo de Gilbert-Elliott parametrizada de acordo com os dados da amostra 4 (b).

Portanto, o resultado da análise da ACF indica que o tempo de duração dos estados B e R não seguem o modelo de Gilbert-Elliott, sendo necessário a investigação de outros parâmetros para a determinação do modelo mais adequado.

4.4 Verificação de distribuição acumulada

Mesmo caracterizada a dependência temporal, foi feita a análise da distribuição acumulada do tempo de duração dos estados B e R, plotando-se a distribuição acumulada complementar do tempo de duração dos estados em escala logarítmica (LLCD), tratado na seção 2.3, conforme mostrado nas Figuras 8 e 9.

Para esta análise, além da plotagem dos dados, foi incluída como referência uma curva com uma distribuição exponencial parametrizada de acordo com cada amostra empírica. Assim, caso o modelo de Gilbert-Elliott se aplicasse ao tempo de duração dos estados B e R das amostras, os gráficos da distribuição exponencial deveriam ficar bastante próximos dos dados amostrados.

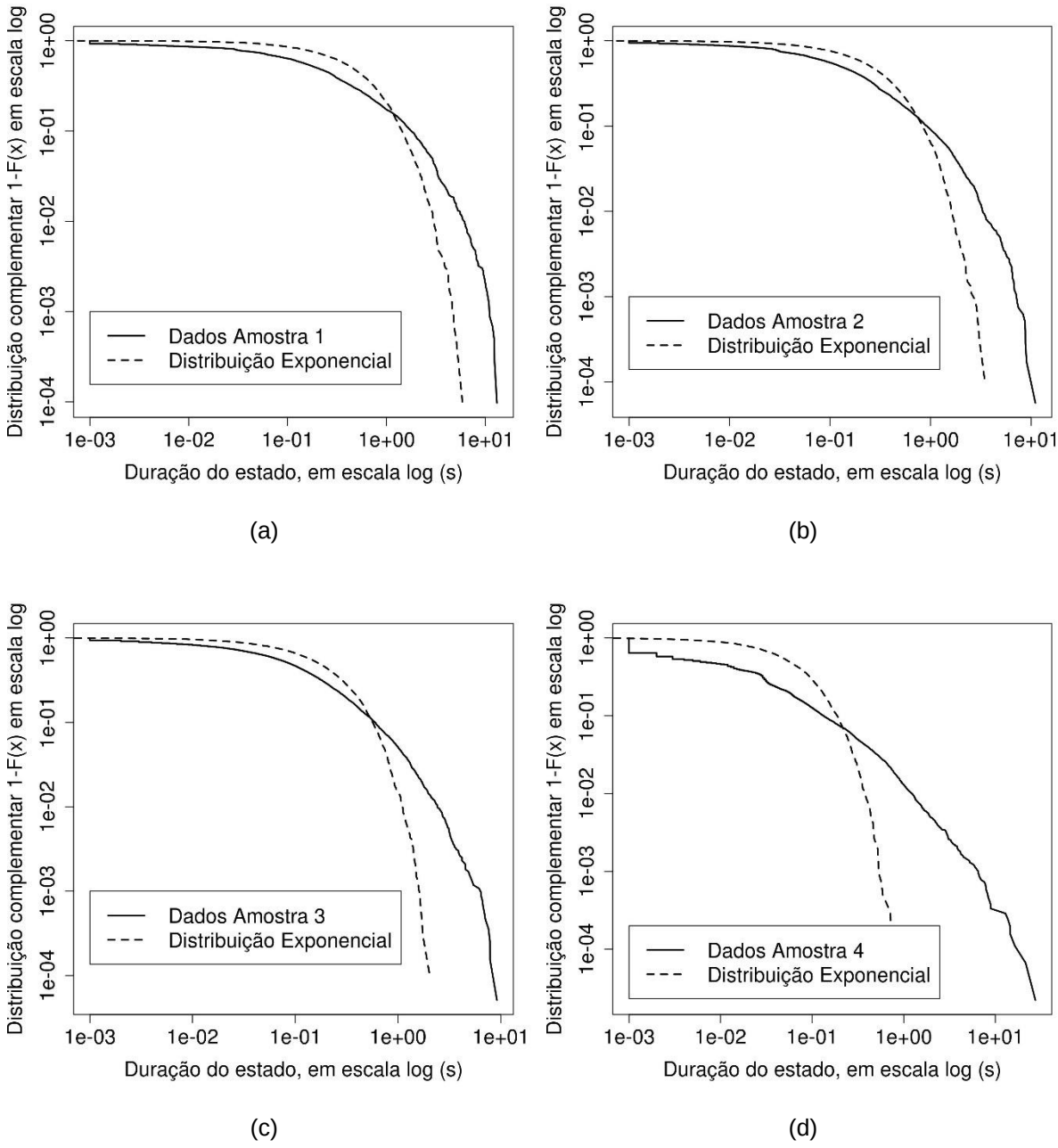


Figura 8 – LLCD do tempo de duração do estado B para a amostra 1 (a), amostra 2 (b), amostra 3 (c) e amostra 4 (d).

Para as quatro amostras percebe-se que a distribuição do tempo de duração do estado B das amostras se difere bastante de uma distribuição exponencial. Para o tempo de duração do estado R fica ainda mais evidente a diferença entre os dados reais e o modelo de Gilbert-Elliott, pois a distribuição fica ainda mais distante da curva da distribuição exponencial.

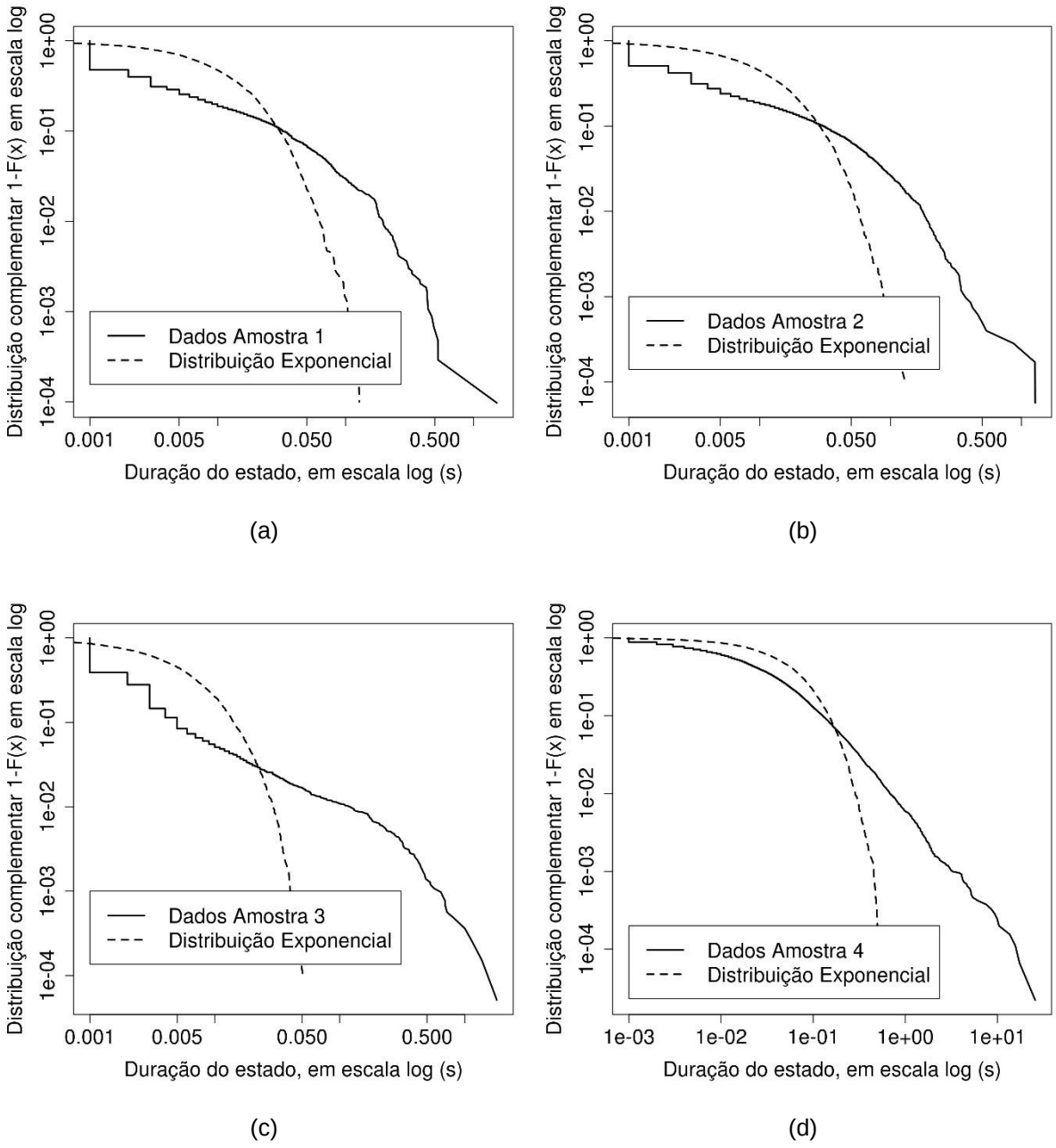


Figura 9 – LLCD do tempo de duração do estado R para a amostra 1 (a), amostra 2 (b), amostra 3 (c) e amostra 4 (d).

Na última amostra identifica-se uma acentuação no efeito de cauda pesada, o que pode estar associado ao fato de que esta amostra apresentou um percentual de perdas bem maior que as anteriores. No entanto, a forte auto correlação observada impede a modelagem utilizando distribuição de cauda pesada diretamente.

Esta distribuição de cauda pesada também já havia sido identificada em uma primeira fase deste trabalho (ROHLING et al., 2017).

4.5 Verificação de estacionariedade

Com a identificação da autocorrelação com decaimento lento nas amostras empíricas, a próxima fase é a análise da estacionariedade das séries. Para verificação da estacionariedade de primeira ordem foi utilizado o teste de Ljung-Box (LJUNG; BOX, 1978), com o software R, que indica se a série é estacionária ou não. O teste desenvolvido por Greta M. Ljung e George E. P. Box é um teste estatístico que verifica se qualquer grupo de autocorrelações em uma série temporal são diferentes de zero. Em vez de testes aleatórios em cada deslocamento distinto, ele testa a aleatoriedade geral, com base em um número definido de deslocamentos. O resultado a ser observado no teste de Ljung-Box é o valor obtido para o parâmetro *p*, que, neste caso, sendo menor do que 0,05 sugere que a série é não estacionária.

Para efeito comparativo com o modelo de Gilbert-Elliott, foi realizado o teste de estacionariedade com a distribuição exponencial gerada a partir da média dos estados BOM e RUIM da quarta amostra, para alguns valores de deslocamento distintos, conforme mostrado na Tabela 7. Com os valores obtidos pode-se observar que a distribuição exponencial apresenta uma estacionariedade para todos os deslocamentos testados, pois a maioria dos valores obtidos ficaram acima de 0,05.

Tabela 7 – Valores do parâmetro p do teste de Ljung-Box para diversos intervalos de uma distribuição exponencial (modelo de Gilbert-Elliott).

Deslocamento (k)	*p* para os estados B	*p* para os estados R
1	0,924	0,9245
5	0,3161	0,3037
10	0,3101	0,4005
15	0,3342	0,3616
20	0,423	0,3925
30	0,3213	0,1898
40	0,3979	0,3395

Aplicando-se o teste de Ljung-Box sobre as séries do tempo de duração dos estados B e R das quatro amostras, foram obtidos os valores mostrados na Tabela 8, utilizando-se os valores de deslocamento para diversos valores entre 1 e 40, sendo obtido sempre o mesmo valor para todas as variações, conforme mostrado na tabela.

Tabela 8 – Valores do parâmetro p do teste de Ljung-Box para as amostras empíricas

Amostra	*p* para o estado B	*p* para o estado R
1	< 2,2e-16	< 2,2e-16
2	< 2,2e-16	< 2,2e-16
3	< 2,2e-16	< 2,2e-16
4	< 2,2e-16	< 2,2e-16

Os valores obtidos sugerem que as séries do tempo de duração dos estados B e R das amostras são não-estacionárias de primeira ordem. Além disso o resultado do teste apresenta uma diferença de comportamento da série em relação ao modelo de Gilbert-Elliott, cujo resultado do teste é apresentado na Tabela 7, diferenciando-se dos resultados obtidos para os dados empíricos.

Para a identificação do modelo, após a caracterização de não estacionariedade de primeira ordem, deve-se verificar então a estacionariedade de segunda ordem.

Uma forma de testar a estacionariedade é a aplicação da transformada discreta de Fourrier (DWIVEDI; SUBBA RAO, 2011), sendo que um dos testes que utiliza este método é o chamado PSR - Priestley-Subba Rao (RAO; SUBBA, 1969). Assim, para verificar a estacionariedade de segunda ordem, foi utilizado o teste PSR, com o software R, que examina um conjunto de densidade espectral verificando o quanto homogêneo é este conjunto com a variação do tempo ou da frequência. A função *stationarity* utilizada gera um resultado com diversos valores numéricos, sendo que o valor chave para verificação da estacionariedade é o valor *p* para a variável *T*, que quanto mais próximo de zero indica uma maior evidência de estacionariedade de segunda ordem. Na Tabela 9 são apresentados os valores obtidos das séries dos estados B e R. Para as quatro amostras do estado B e R o valor de *p* para a variável *T* obtido foi igual a zero, o que indica uma estacionariedade de segunda ordem para as quatro amostras.

Tabela 9 – Valores do teste de estacionariedade de segunda ordem (PSR) para as amostras empíricas do tempo de duração dos estados B e R

	p-value for T			
	Amostra 1	Amostra 2	Amostra 3	Amostra 4
Estados B	0	0	0	0
Estados R	0	0	0	0

Portando os testes de estacionariedade indicam que os dados empíricos não possuem estacionariredade de primeira ordem, mas possuem estacionariedade de segunda ordem, tanto para o tempo de duração do estado B quanto para o estado R.

Assim, o modelo a ser investigado deverá apresentar estacionariedade de segunda ordem, sem estacionarieade de primeira ordem, sendo um dos mais utilizados para isto o modelo FARIMA.

4.6 Duração dos estados

Para ilustrar o comportamento da série foram plotados os gráficos com o tempo de duração dos estados B e R, mostrados nas Figuras 10 e 11. Uma das diferenças identificada entre os gráficos foi a quantidade de mudanças de estado na quarta amostra, com mais de 20.000 estados, consequência da maior perda de pacotes ocorrida durante a realização desta amostra, levando a uma maior troca de estados durante o período do teste. Também como consequência do maior percentual de perdas, observa-se graficamente um menor valor médio do tempo de duração do estado B, bem como uma maior variação na amplitude dos gráficos, com valores máximos maiores, fato já identificado na acentuação do efeito de cauda pesada no gráfico da distribuição.

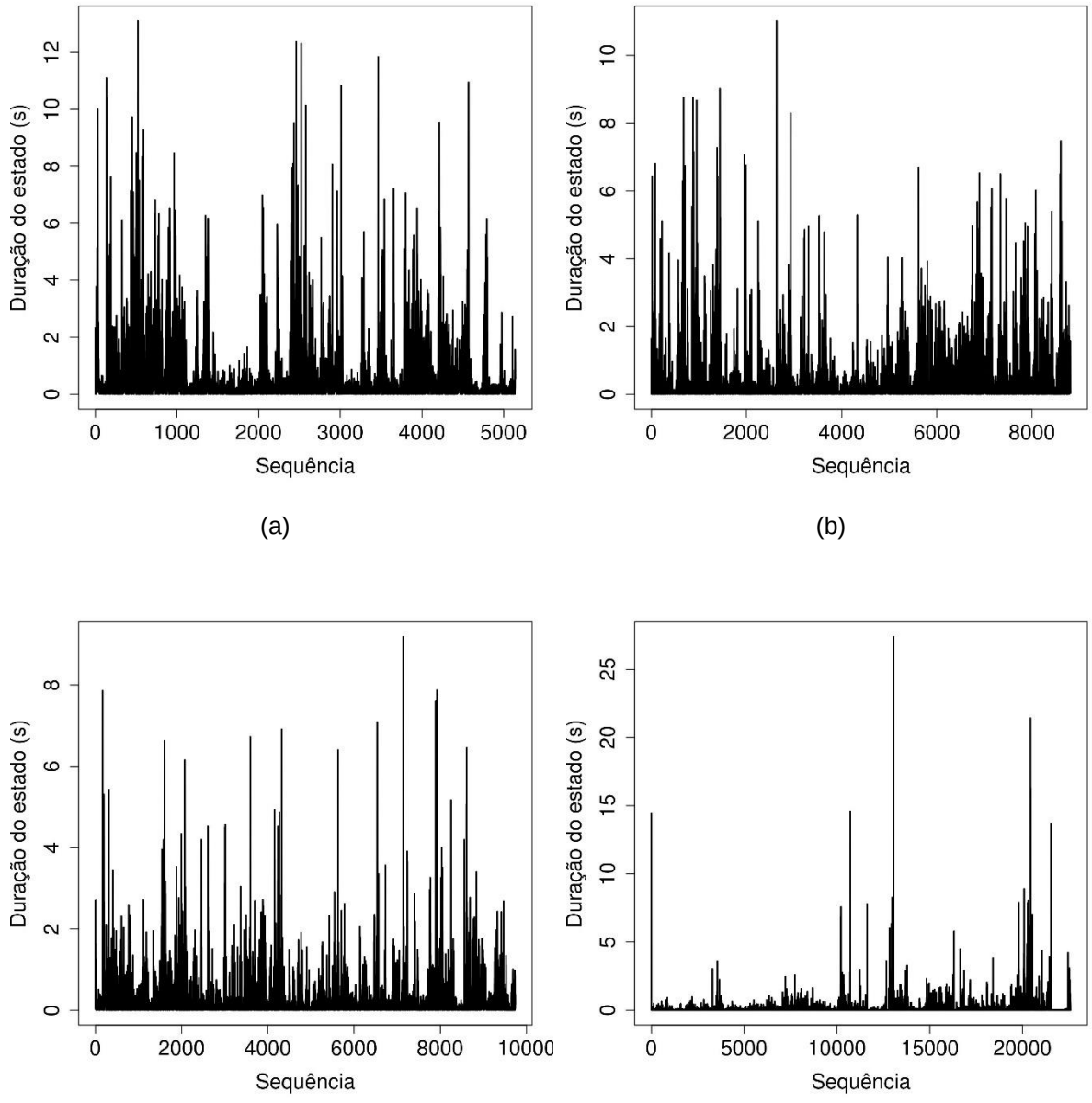


Figura 10 – Séries do tempo de duração do estado B da amostra 1 (a), amostra 2 (b), amostra 3 (c) e amostra 4 (d)

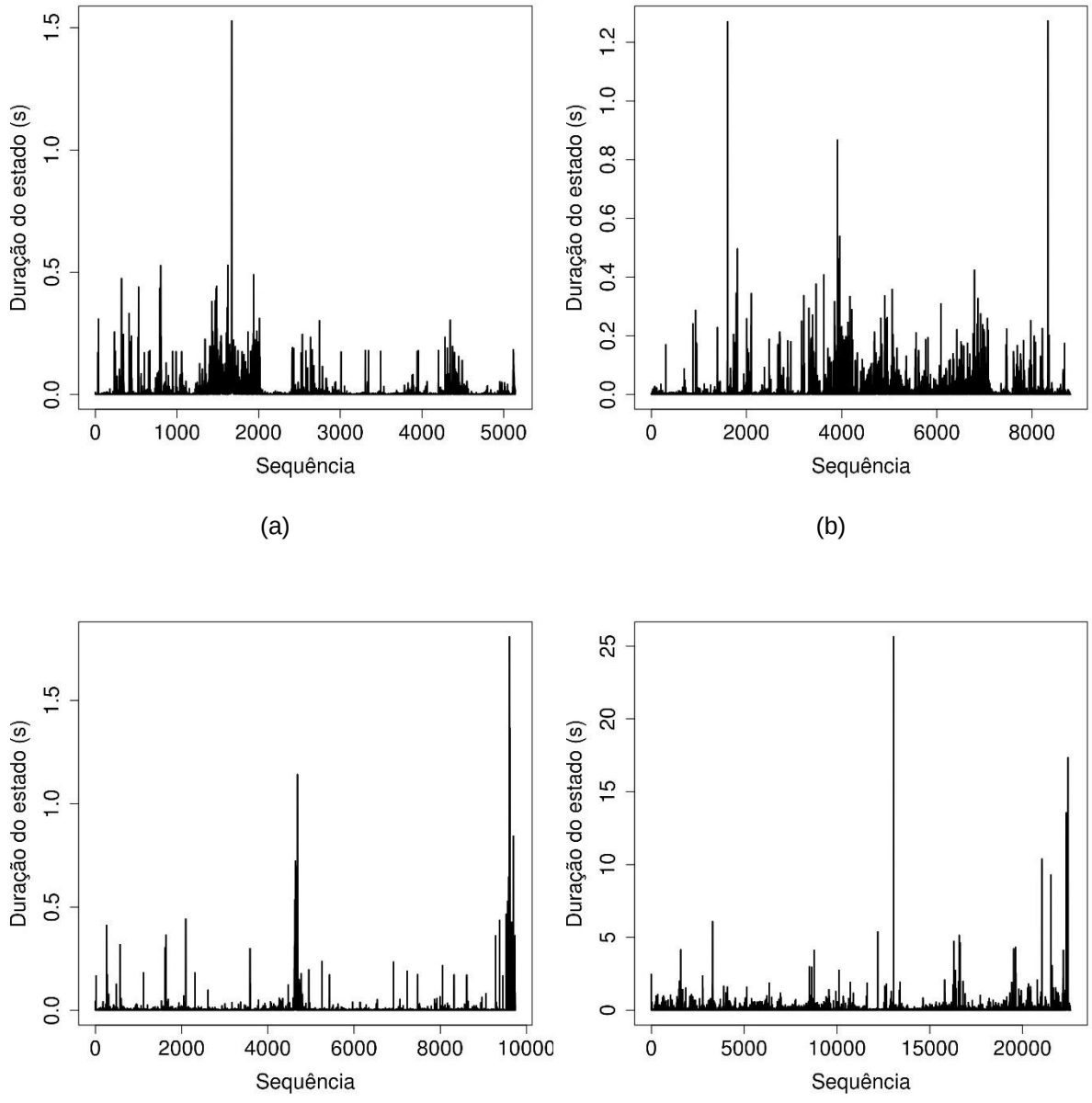


Figura 11 – Séries do tempo de duração do estado R da amostra 1 (a), amostra 2 (b), amostra 3 (c) e amostra 4 (d)

4.7 Possíveis modelos

Como os dados empíricos apresentaram uma característica de memória de longa duração, pela análise da ACF, os modelos que não possuem esta característica podem ser descartados como possíveis modelos para o nosso estudo, tal como as Cadeias de Markov, utilizadas no modelo de Gilbert-Elliott. Pela característica de estacionariedade das séries do estado B e R, também o modelo ARMA não é um modelo adequado para estas séries. E a dependência temporal de longa duração, indentificada na ACF também leva a descartar o modelo ARIMA como possível para modelar os dados empíricos.

Portanto, para este estudo, o modelo a ser investigado deve caracterizar séries estacionárias, pois o resultado do teste de estacionariedade indica se tratar de uma série estacionária de segunda ordem, e com memória de longa duração, em função do resultado observado na ACF com decaimento lento. Alguns dos modelos que atendem estes requisitos são o FARIMA, fBm, On-Off de cauda pesada, sendo escolhido para a análise o modelo FARIMA, descrito no item 2.7.

Apenas para ilustração foi realizada a modelagem da série utilizando-se o modelo ARMA a partir dos dados da quarta amostra. O melhor ajuste obtido foi para ARMA(2,0) e os valores obtidos para os parâmetros $\phi 1$ e $\phi 2$ são mostrados na Tabela 10.

Tabela 10 – Valores dos parâmetros $\phi 1$ e $\phi 2$ para a quarta amostra empírica do tempo de duração dos estados B e R

	$\phi 1$	$\phi 2$
Estados B	0,164408	0,121494
Estados R	0,309247	0,041007

Após a obtenção dos parâmetros, foi gerada a série com o modelo ARMA, com previsão de um passo, sendo elaborado o gráfico a partir de um intervalo dos valores do tempo de duração do estado B e R da quarta amostra, comparando-se os dados empíricos com o modelo ARMA, conforme mostrado na Figura 12. Observa-se que o modelo ARMA não consegue modelar adequadamente os dados empíricos, não conseguindo acompanhar a variação e muito menos os valores de pico da série. Isto

já seria esperado pois a dependência temporal identificada nos gráficos de ACF são de longa duração, e os modelos usuais AR, MA ou ARIMA não possuem esta característica. Deste modo, deve ser investigado então o modelo FARIMA.

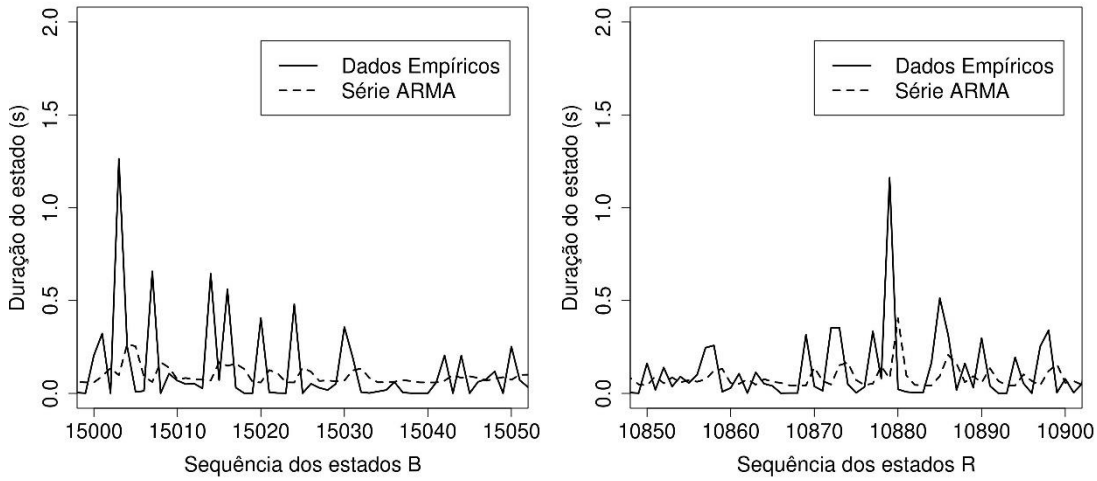


Figura 12 – Comparação dos empíricos, do tempo de duração do estado B e R da amostra 4, com o modelo ARMA.

Para a investigação do modelo FARIMA foi utilizado o Sofware R com as funções do pacote fArma (WUERTZ, 2015). Na primeira fase foram gerados os valores dos parâmetros *p*, *q* e *d* para a amostra 4, cujos resultados obtidos são mostrados nas Tabelas 11 e 12.

Tabela 11 – Estimação dos parâmetros d, p e q para o tempo de duração dos estado B da quarta amostra.

(p,q)	d	$\phi 1$	$\phi 2$	$\theta 1$
(1,0)	0,212924	-0,084853		
(1,1)	0,30390	0,41718		0,59434
(2,0)	0,223904	- 0,102744	-0,037959	

Tabela 12 – Estimação dos parâmetros d, p e q para o tempo de duração dos estado R da quarta amostra.

(p,q)	d	$\phi 1$	$\phi 2$	$\theta 1$
(1,0)	0,099349	0,208032		
(1,1)	0,09204	0,25368		0,03958
(2,0)	0,091320	0,214774	0,009260	

Como o modelo apresentou a menor probabilidade de erro na estimação dos parâmetros para a combinação (2,0), foi adotado este padrão para as quatro amostras.

No método utilizado são calculados os estimadores para os parâmetros do modelo FARIMA (p, d, q) usando o método de (HASLETT J. AND RAFTERY A.E., 1989). Os coeficientes obtidos para a duração do tempo dos estados B e R, das quatro amostras, são mostrados nas Tabelas 13 e 14.

Tabela 13 – Parâmetros da função FARIMA para a duração do estado B

Amostra	(p,q)	d	$\phi 1$	$\phi 2$
1	(2,0)	0,3407	-0,2421	-0,07109
2	(2,0)	0,2695	-0,1227	-0,04981
3	(2,0)	0,3123	-0,1641	-0,06363
4	(2,0)	0,22390	-0,10274	-0,03796

Tabela 14 – Parâmetros da função FARIMA para a duração do estado R

Amostra	(p,q)	d	$\phi 1$	$\phi 2$
1	(2,0)	0,2891	-0,2394	-0,09797
2	(2,0)	0,2237	-0,1401	-0,06581
3	(2,0)	0,3089	-0,2421	-0,07436
4	(2,0)	0,09132	0,21477	0,00926

Para uma inspeção visual de aderência foram geradas as séries simuladas com o modelo FARIMA, com os parâmetros correspondentes às séries empíricas, mostradas nas Figuras 12 e 13.

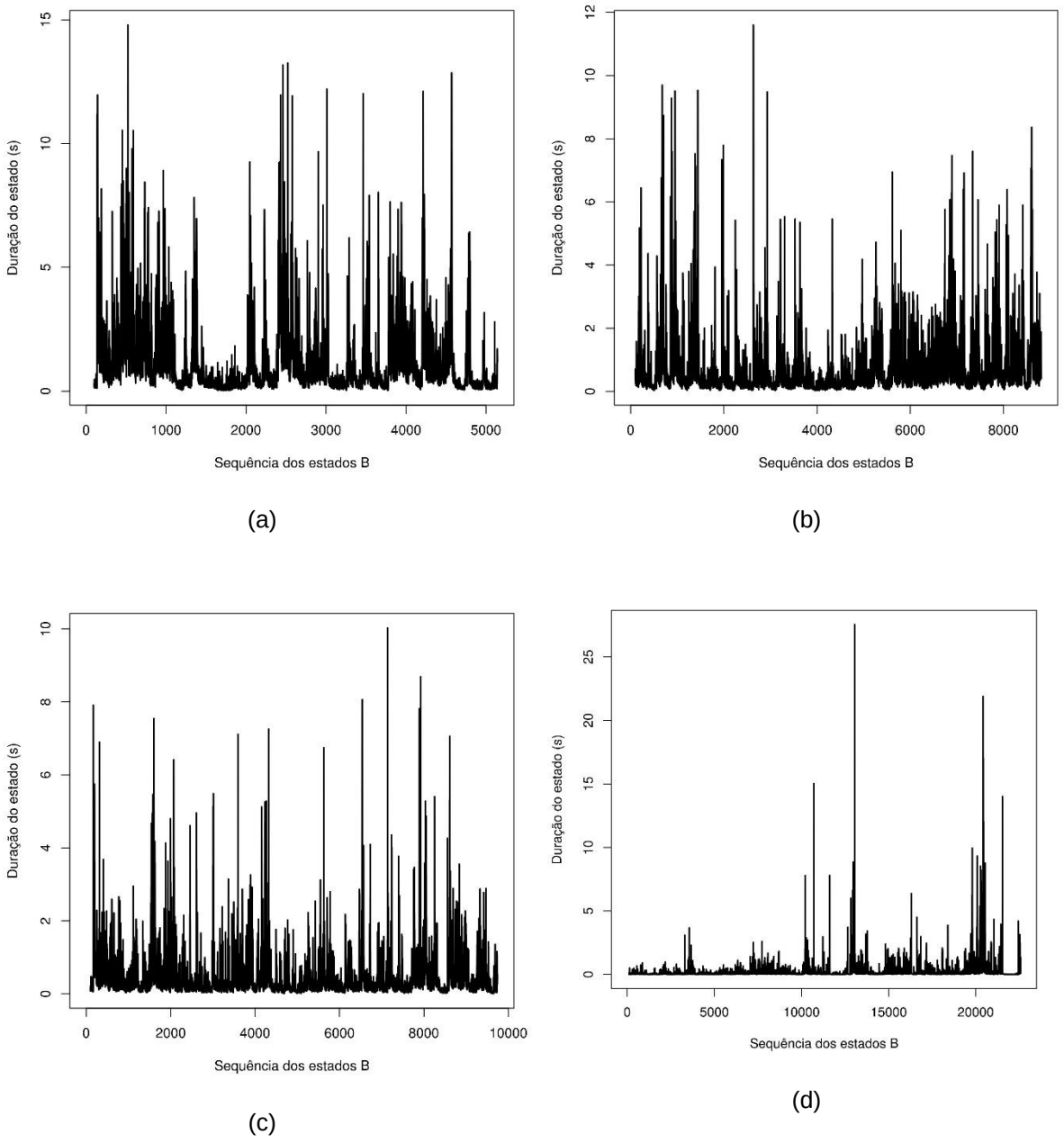


Figura 13 – Modelo FARIMA para o tempo de duração dos estados B da rede para a amostra 1 (a), amostra 2 (b), amostra 3 (c) e amostra 4 (d)

Comparando-se os dados obtidos pela série FARIMA, apresentados na Figura 13, com a duração dos estados B, apresentado na Figura 10, observa-se que o comportamento dos dados empíricos é semelhante à série sintética gerada pelo FARIMA.

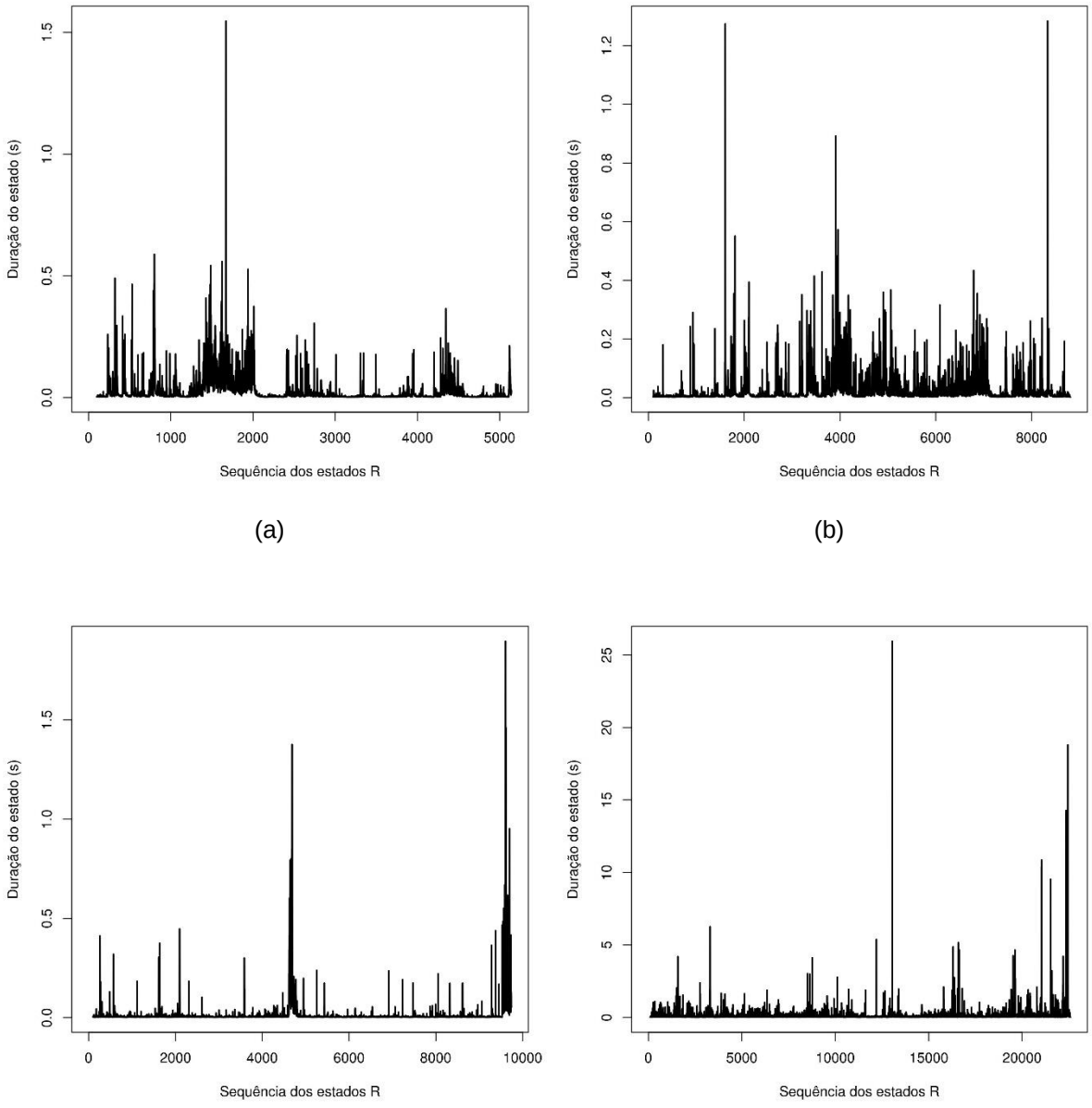


Figura 14 – Modelo FARIMA para o tempo de duração dos estados R da rede para a amostra 1 (a), amostra 2 (b), amostra 3 (c) e amostra 4 (d)

Comparando-se os dados obtidos pela série FARIMA, apresentados na Figura 14, com a duração dos estados R, apresentado na Figura 11, observa-se que também neste caso a série sintética gerada pelo FARIMA consegue acompanhar o comportamento dos dados empíricos.

4.8 Análise do Resíduo

A partir das séries geradas pelo modelo FARIMA, foram geradas as séries contendo o resíduo, obtido pela diferença entre os valores das séries empíricas e os valores do modelo. Para avaliar o resultado do modelo foi utilizado o gráfico Quantile-Quantile (QQPlot) descrito em (BECKER et al., 1988). O QQPlot é uma ferramenta gráfica utilizada para comparar características de duas populações. Nesta técnica o conjunto de dados é ordenado em ordem de grandeza, sendo os valores que dividem o conjunto em quatro partes iguais, chamados quartis, em dez partes os decis, e em N partes, que podem corresponder ao número de dados do conjunto, são denominados de quantis. Neste gráfico os pontos representam os quantis de cada uma das amostras, colocados nos eixos *x* e *y*. Se duas amostras vêm da mesma população, os pontos devem estar em torno de uma linha diagonal em 45º sobre a origem. Comparando-se os pontos traçados no mesmo gráfico com esta linha diagonal, caso os pontos estejam em uma linha paralela à diagonal, as duas distribuições possuem distribuição semelhante e um processo está localizado em um nível mais alto em relação ao outro. Neste trabalho foi empregado o software estatístico R para construção dos gráficos, com a utilização da função QQPlot, que traça ainda uma linha de comparação entre os quantiles da distribuição empírica e os quantiles de uma distribuição calculada.

Como o resíduo deveria apresentar uma distribuição normal, foi traçado o QQPlot comparando-se o resíduo com uma distribuição normal, gerada a partir do valor médio e desvio padrão de cada uma das quatro amostras, cujo resultado pode ser visto na Figura 15, para o resíduo da duração do estado B, e na Figura 16 para o resíduo da duração do estado R.

A Figura 15 apresenta o QQPlot comparando o resíduo na previsão de um passo para o tempo de duração do estado B. O gráfico apresenta no eixo horizontal os quantiles teóricos da distribuição normal e no eixo vertical os valores do resíduo. A aderência perfeita é denotada pela linha contínua no gráfico. Observa-se que existe uma diferença significativa, o que indica que o FARIMA não capturou todas as características do estado B.

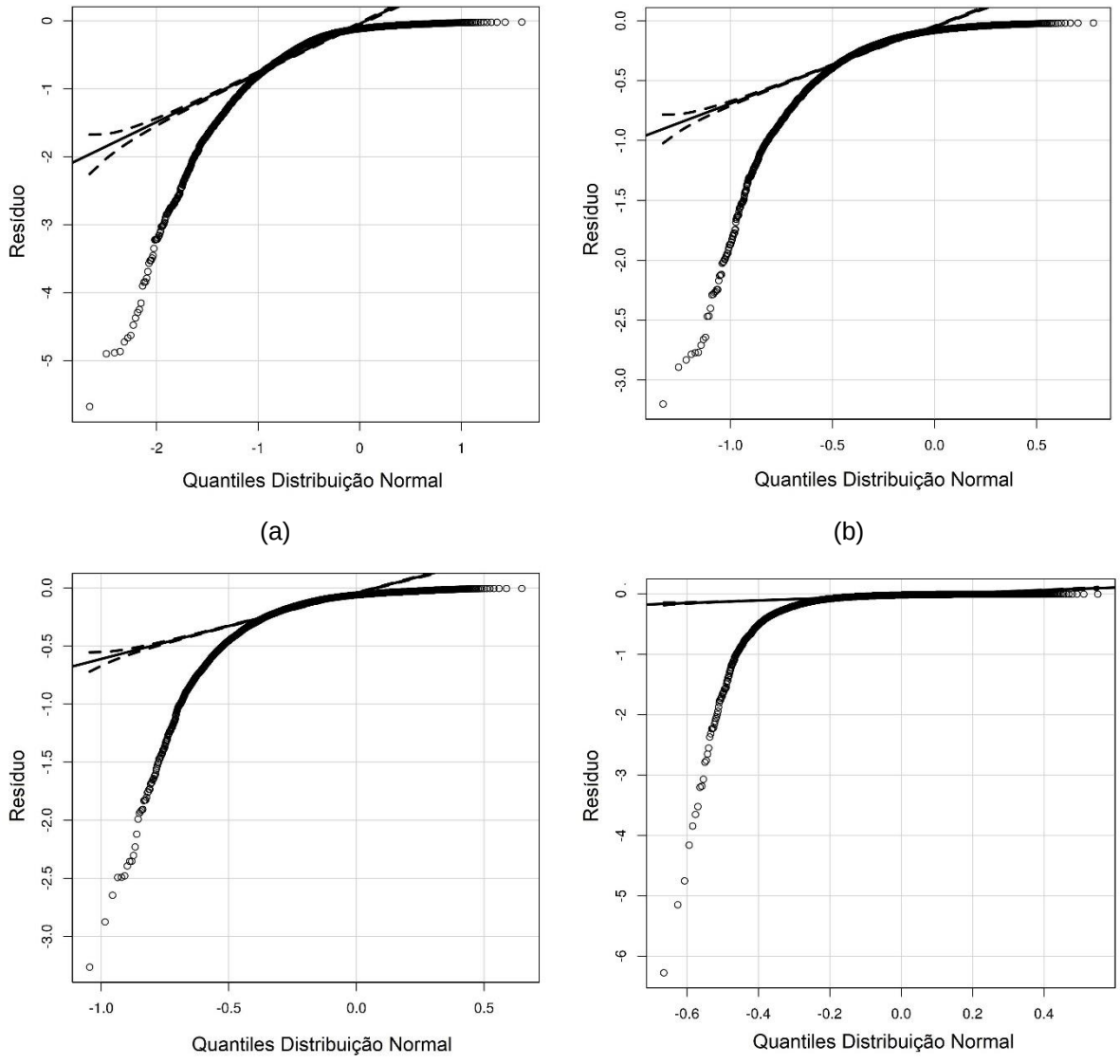


Figura 15 – QQplot do resíduo do modelo FARIMA com uma distribuição normal, para o tempo de duração dos estados B da rede para a amostra 1 (a), amostra 2 (b), amostra 3 (c) e amostra 4 (d)

A Figura 16 apresenta o QQPlot comparando o resíduo na previsão de um passo para o tempo de duração do estado R utilizando o FARIMA com a distribuição normal. O gráfico apresenta no eixo horizontal os quantiles teóricos da distribuição normal e no eixo vertical os valores do resíduo. A aderência perfeita é denotada pela linha contínua no gráfico. Observa-se que existe uma diferença significativa, o que indica que o FARIMA não capturou todas as características do estado R.

Portanto, apenas o modelo FARIMA não é suficiente para modelar os dados empíricos.

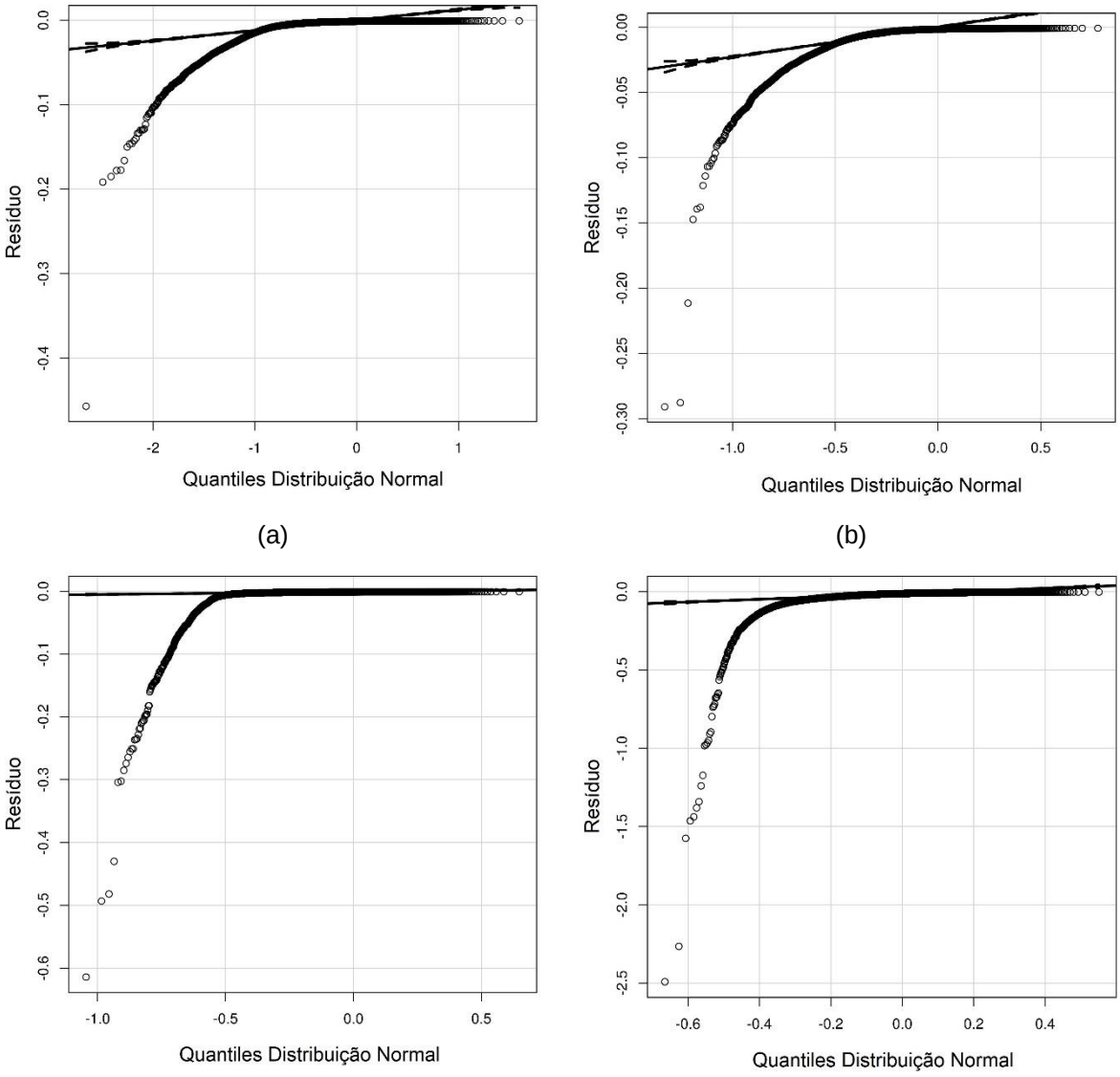


Figura 16 – QQplot do resíduo do modelo FARIMA com uma distribuição normal, para o tempo de duração dos estados R da rede para a amostra 1 (a), amostra 2 (b), amostra 3 (c) e amostra 4 (d)

Outra análise do resíduo é a observação da existência de uma dependência temporal, que é realizada pela elaboração do gráfico da ACF para o resíduo do modelo FARIMA, para a duração dos estados B e R, mostrados nas Figuras 17 e 18.

Caso o resíduo das amostras seja apenas o ruído branco, que é representado por uma distribuição normal, a ACF deve apresentar valores próximos à zero para todos os deslocamentos diferentes de zero.

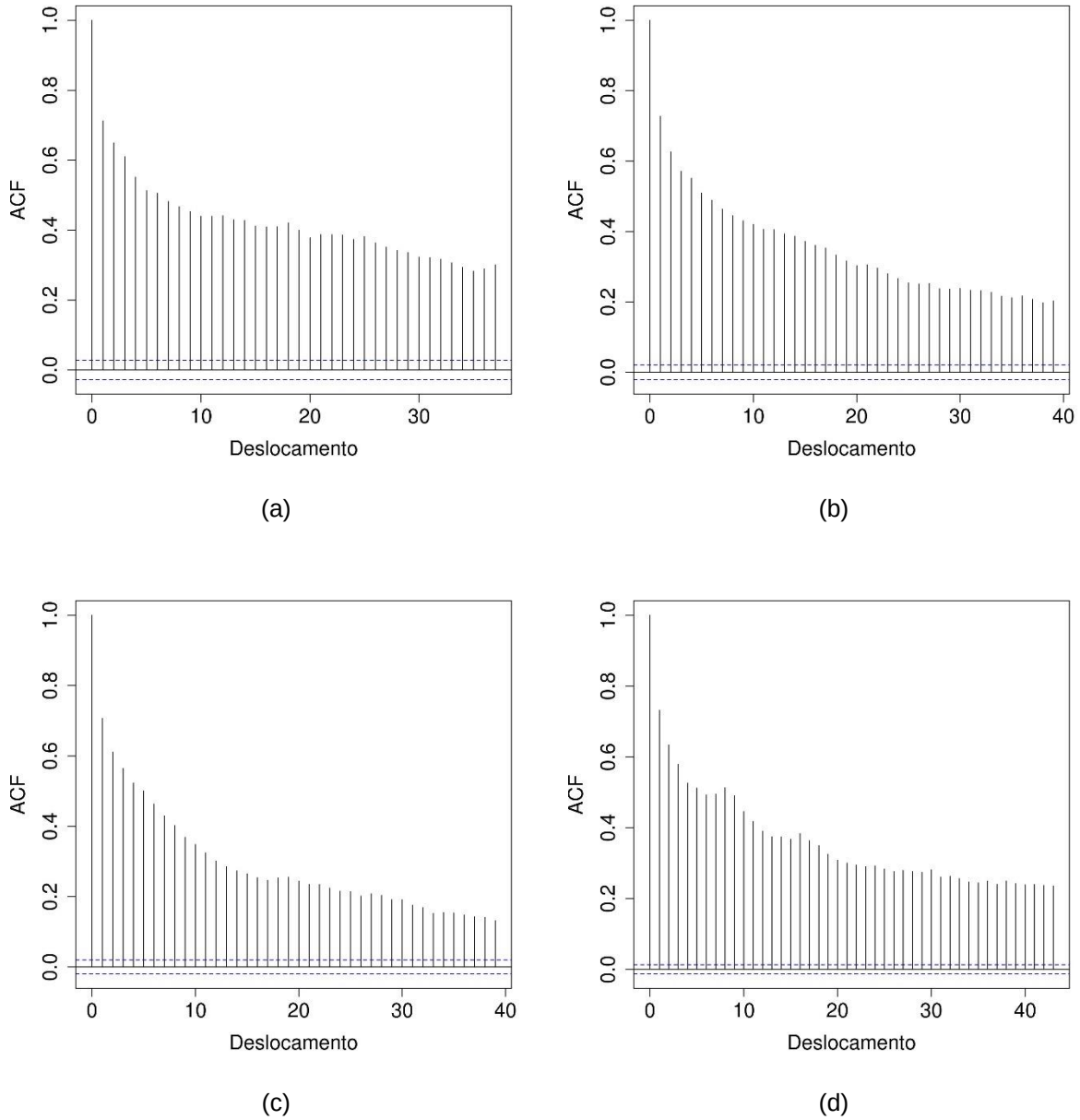
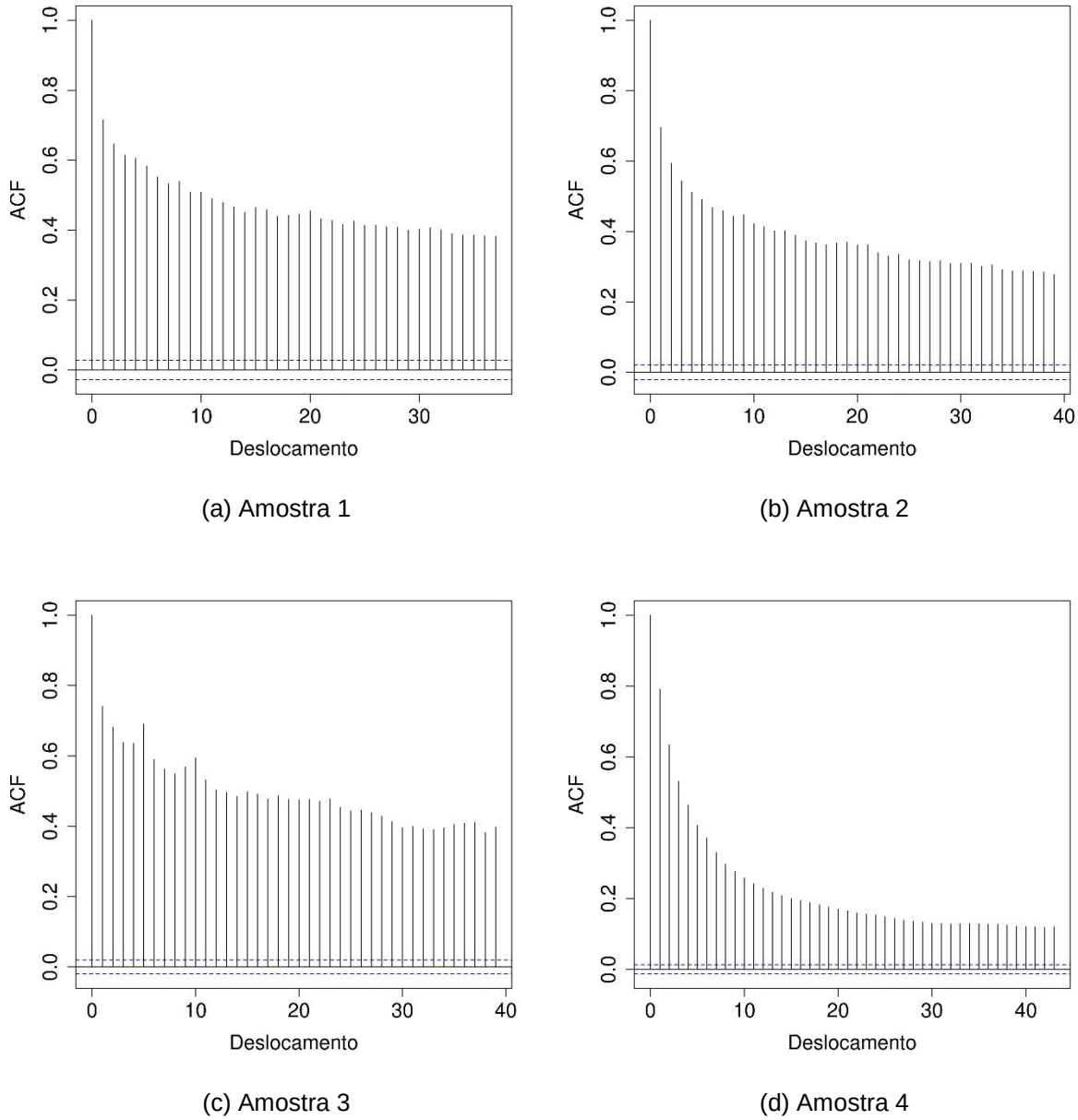


Figura 17 – ACFdo resíduo do modelo FARIMA com uma distribuição normal, para o tempo de duração dos estados B da rede para a amostra 1 (a), amostra 2 (b), amostra 3 (c) e amostra 4 (d)

Para as quatro amostras do resíduo do modelo FARIMA, para o tempo de duração do estado B, observa-se a existência de uma dependência temporal de longa duração, o que indica que o resíduo não é apenas um ruído branco.



(a) Amostra 1

(b) Amostra 2

(c) Amostra 3

(d) Amostra 4

Figura 18 – ACFdo resíduo do modelo FARIMA com uma distribuição normal, para o tempo de duração dos estados R da rede.

Para as quatro amostras do resíduo do modelo FARIMA, para o tempo de duração do estado R, observa-se também uma dependência temporal de longa duração, indicando que neste caso o resíduo também não é apenas um ruído branco.

Portanto, apesar do QQPlot do resíduo do tempo de duração do estado R ter apresentado uma maior linearidade, a análise da ACF indica que o modelo FARIMA possui um resíduo que contém ainda parte dos dados empíricos, não sendo apenas um ruído branco.

Na sequência busca-se compreender melhor as causas da falha de aderência do modelo FARIMA aos dados empíricos. Foi tomado um intervalo da amostra 4 contendo variações do tempo de duração dos estados B e R, mostrado na Figura 15, e produzida uma previsão de um passo com o modelo FARIMA, sendo o resultado apresentado pela linha tracejada na Figura 19. Nesta figura observa-se que o modelo FARIMA acompanha a variação de subida, apresentando um atraso na descida. Observa-se que o FARIMA consegue acompanhar todas as variações, representando de maneira muito mais próxima a série empírica do que o modelo ARMA mostrado na Figura 12.

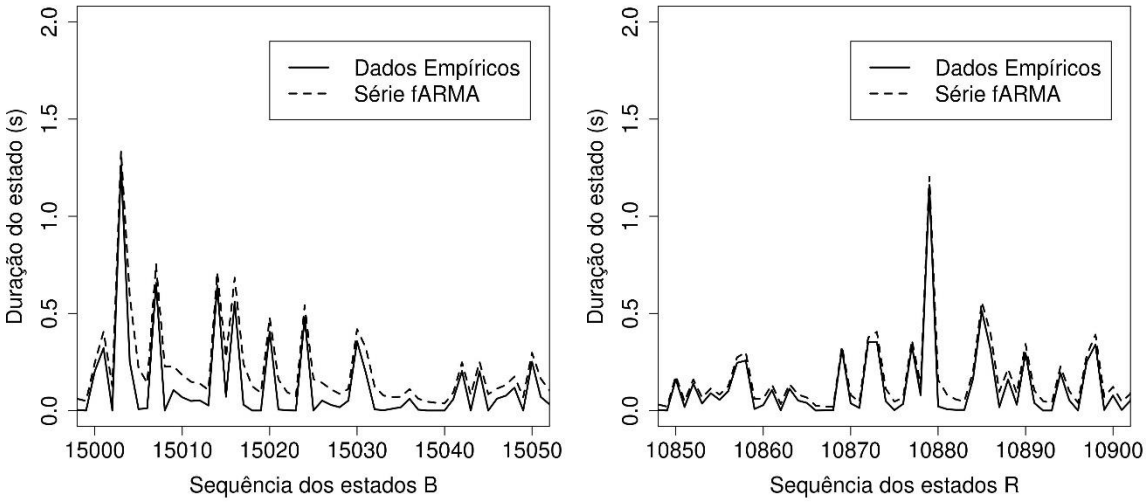


Figura 19 – Comparativo para o tempo de duração dos estados B e R da amostra 4.

Desta forma, é necessário a criação de um modelo complementar, a ser usado em conjunto com o modelo FARIMA, para conseguir representar a série dos dados empíricos com maior aderência.

5 CONCLUSÕES E TRABALHOS FUTUROS

Para caracterizar os surtos de erros na rede WiFi foi aplicada uma técnica de clusterização utilizando o método k-means. A base de dados obtida, com a duração dos estados B e R da rede em estudo, foi devidamente preparada para a análise estatística, visando a obtenção do modelo que melhor represente perdas de pacotes em redes WiFi.

Feita a análise das séries do tempo de duração dos estados B e R, identificou-se uma dependência temporal, o que indica que a perda de pacotes não segue o modelo clássico de Gilbert-Elliott. E esta hipótese também é reforçada pela análise dos gráficos de distribuição acumulada complementar.

O comportamento observado nas séries empíricas aponta para um modelo com dependência temporal, de longa duração e estacionário de 2ª ordem. Com estas características, os modelos clássicos não seriam adequados, sendo investigado neste trabalho o modelo FARIMA. A análise do ruído indica que o modelo FARIMA não adere perfeitamente aos dados empíricos. No entanto, uma análise mais cuidadosa revela que o modelo FARIMA está muito próximo, sendo este um ponto de partida promissor para o desenvolvimento do novo modelo.

Esta dissertação é parte de um projeto maior, que envolve uma tese de doutorado. Nesta tese de doutorado será realizado o desenvolvimento final do modelo de perdas, utilizando o conhecimento gerado por esta dissertação.

Como continuação deste trabalho, cabem trabalhos futuros desenvolvendo o modelo apropriado para representar os dados empíricos observados. Outro trabalho importante será o estudo das perdas que ocorrem dentro de cada estado B e R, com a respectiva modelagem. Podem ser aprofundadas as discussões a respeito da clusterização, com o uso de mais de dois estados, de acordo com o indicador no dendograma. Também é possível trocar o algoritmo de clusterização para confirmar os resultados. Assim, pode-se buscar uma modelagem mais adequada para a perda de pacotes em redes WiFi, considerando-se a abstração do modelo de Gilbert-Elliott de estados B e R da rede.

REFERÊNCIAS

AL-FUQAHA, A.; GUIZANI, M.; MOHAMMADI, M.; ALEDHARI, M.; AYYASH, M., Internet of Things: A Survey on Enabling Technologies, Protocols and Applications. **IEEE Communications Surveys & Tutorials**, v. 17, n. 4, p. 2347–2376, 2015.

ANGEJA, J.; NAVARRO, A., A New Packet loss Model of the IEEE 802.11 g Wireless Network for Multimedia Communications. **Consumer Electronics, IEEE**, p. 809–814, 2005.

BABU, C. N.; REDDY, B. E., A Moving-average Filter Based Hybrid ARIMA – ANN Model for Forecasting Time Series Data. **Applied Soft Computing Journal**, v. 23, p. 27–38, 2014. Elsevier B.V.

BECKER, R. A.; CHAMBERS, J. M.; WILKS, A. R., The New S Language. **Wadsworth & Brooks/Cole**, 1988.

BOX, G. E. P.; JENKINS, G., Time Series Analysis, Forecasting and Control. **Holden-Day, Incorporated**, 1990.

DWIVEDI, Y.; SUBBA RAO, S., A Test for Second-order Stationarity of a Time Series Based on the Discrete Fourier Transform. **Journal of Time Series Analysis**, v. 32, n. 1, p. 68–91, 2011. Blackwell Publishing Ltd.

ELLIOTT, E., Estimates of Error Rates for Codes on Burst-Noise Channels. **Bell System Technical Journal**, v. 42, p. 1977–1997, 1963.

GILBERT, E. N. Capacity of a Burst-Noise Channel. **Bell System Technical Journal**, v. 39, p. 1253–1265, 1960.

GRINSTEAD, C. M.; SNELL, J. L., Introduction to Probability. **American Mathematical Society**, 2007.

HASLETT J. AND RAFTERY A.E., Space-time Modelling with Long-memory Dependence. **Applied Statistics**, p. 1–50, 1989.

HASSLINGER, G.; HOHLFELD, O., The Gilbert-Elliott Model for Packet Loss in Real Time Services on the Internet. **Measuring, Modelling and Evaluation of Computer and Communication Systems (MMB), 2008 14th GI/ITG Conference**, p. 1–15, 2008.

HAYKIN, S.; VEEN, B. VAN. Signal and Systems. **John Wiley & Sons**, 1999.

IEEE, Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications. **IEEE Standard 802.11-1997**. The Institute of Electrical and Electronics Engineers, 1997.

ITU-T, I. T. U. G.1050 - Network Model for Evaluating Multimedia Transmission Performance over Internet Protocol, **International Telecommunication Union**, 2011.

KARAGIANNIS, T.; MOLLE, M.; FALOUTSOS, M., Long-range Dependence Ten Years of Internet Traffic Modeling. **IEEE Internet Computing**, v. 8, n. 5, p. 57–64, 2004.

KRUNZ, M. M.; KIM, J. G., Fluid Analysis of Delay and Packet Discard Performance for QoS Support in Wireless Networks. **IEEE Journal on Selected Areas in Communications**, v. 19, n. 2, p. 384–395, 2001.

LEE, K. K.; CHANSON, S. T., Packet Loss Probability for Real-time Wireless Communications. **IEEE Transactions on Vehicular Technology**, v. 51, n. 6, p. 1569–1575, 2002.

LJUNG, G. M.; BOX, G. E. P., On a Measure of Lack of Fit in Time Series Models. **Biometrika**, p. 297–303, 1978.

MEYER, P. L. Probabilidad Aplicações à Estatística. **LTC - Livros Técnicos e Científicos Editora S.A**, 1983

PARK, K.; WILLINGER, W., Self-similar Network Traffic and Performance Evaluation. **Wiley-Interscience**, 2000.

RAO, M. B. P.; SUBBA, T., A Test for Non-Stationarity of Time-Series. **Journal of the Royal Statistical Society . Series B (Methodological)**, v. 31, n. 1, p. 140–149, 1969.

ROHLING, L. J.; SILVA, C. A. G.; FERNÁNDEZ, E. M. G.; PEDROSO, C. M., Evidências de Falha no Modelo de Gilbert-Elliott. **XXXIV Simpósio Brasileiro de Telecomunicações – SBrT2016**, 2016.

RUSS, S. H.; HAGHANI, S., 802.11g Packet-Loss Behavior at High Sustained Bit Rates in the Home. **IEEE Transactions on Consumer Electronics**, v. 55, n. 2, p. 788–791, 2009.

SINGH, A.; RANA, A.; PRADESH, U. K-means with Three Different Distance Metrics. **International Journal of Computer Applications**, v. 67, n. 10, p. 13–17, 2013.

VELLA, J.; ZAMMIT, S., Packet Losses of Multicast over 802.11n Heterogeneous Wireless Local Area Network. **Strategic Educational Pathways Scholarship**, April 2012, p. 11–13, 2013.

VIEIRA CARDOSO, K.; REZENDE, J. F., Accurate Hidden Markov Modeling of Packet Losses in Indoor 802.11 Networks. **IEEE Communications Letters**, v. 13, n. 6, p. 417–419, 2009.

WITTEN, I. H.; FRANK, E.; HALL, M. A., Data Mining: Practical Machine Learning **Tools and Techniques,** Third Edition, 2011.

WUERTZ, D., Package “ fArma .” **http://www.rmetrics.org**, 2015.

YAO, Q.; BROCKWELL, P. J., Gaussian Maximum Likelihood Estimation for ARMA Models II: Spatial Processes. **Bernoulli**, v. 12, n. 3, p. 403–429, 2006.

YOUNESIAN, E.; KHALEEL, H.; DELGADO, M. T.,. Packet-Loss Modelling for Multi-Radio Wireless Sensor Networks. **IEEE 10th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob),** p. 673–678, 2014.

YU, X.; MODESTINO, J. W.; TIAN, X., The Accuracy of Markov Chain Models in Predicting Packet-loss Statistics for a Single Multiplexer. **IEEE Transactions on Information Theory**, v. 54, n. 1, p. 489–501, 2008.